Huawei Enterprise A Better Way

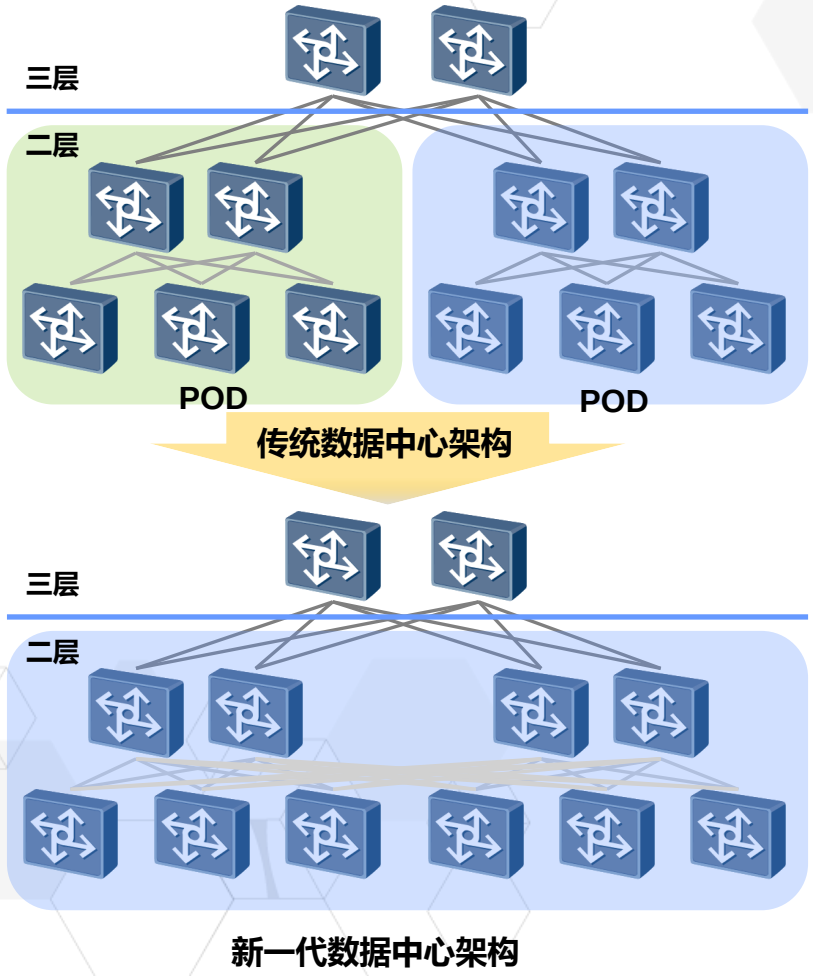
数据中心网络TRILL新特性

www.huawei.com

HUAWEI

目录

TRILL概述—数据中心发展趋势



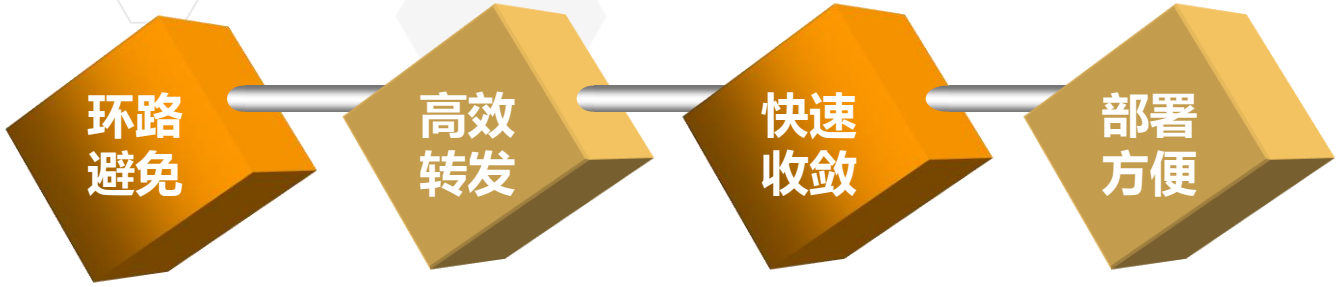
传统数据中心架构

- 传统数据中心组网方式，一般二层只到接入或汇聚交换机，虚拟机的迁移只能局限一个二层区域内。如果需要跨二层区域迁移，需要更改VM的IP地址，如果没有负载分担LoadBalance屏蔽等手段，应用会中断。

新一代数据中心架构

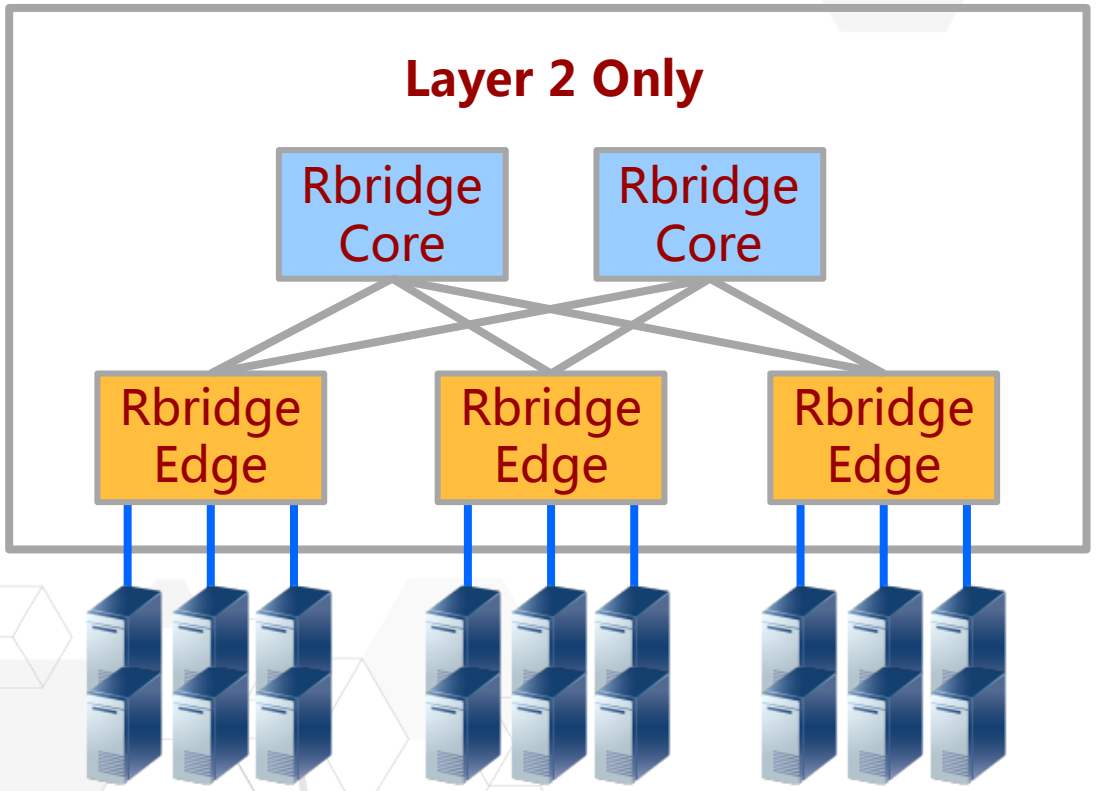
- 在云计算时代，IDC运营商为了更充分的利用数据中心资源，VM需要更大的迁移范围；
- 由于服务器之间存在大量的横向流量，要求数据报文支持无阻塞转发，网络链路资源得到充分的利用。

TRILL概述—TRILL优点



构建无环分发树及引入TTL避免环路。	数据流量基于SPF及ECMP快速转发。	网络变化实时侦听全网拓扑亚秒收敛。	配置简单单播及组播业务单控制协议。

TRILL概述—TRILL概念



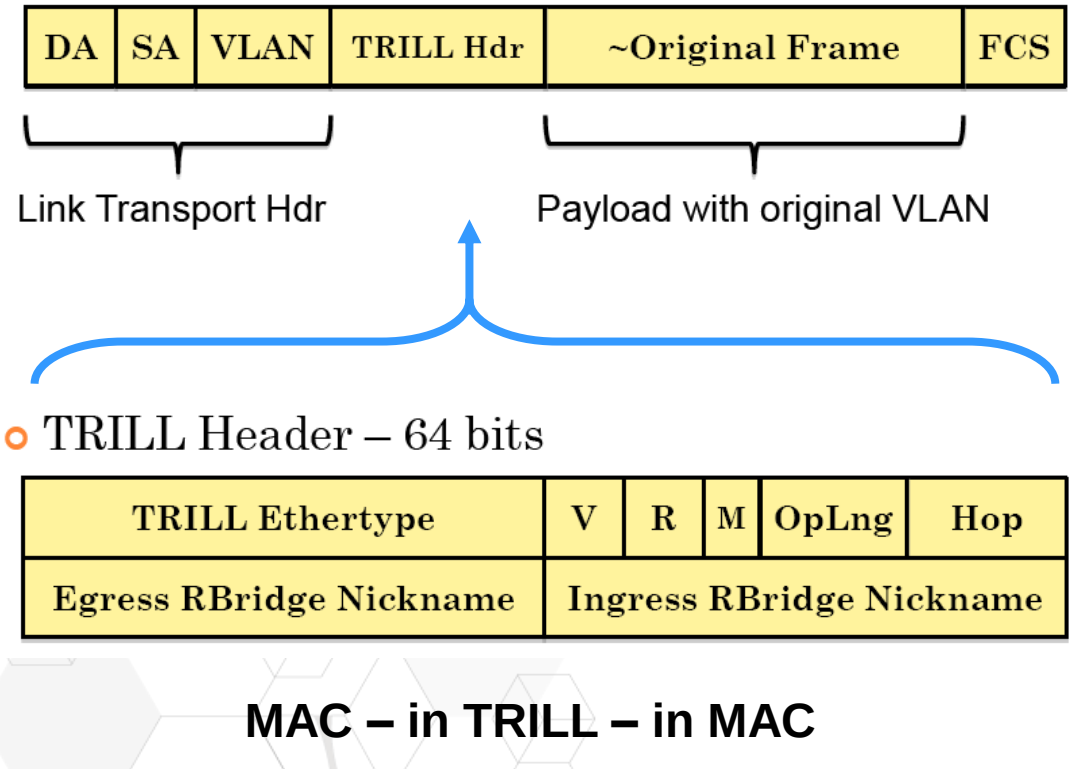
TRILL概念

TRILL(Transparent Interconnection of Lots of Links)是一种在二层网络上基于链路状态计算的路由协议，它通过扩展IS-IS协议来实现，运行TRILL协议的设备叫做RB（Route Bridge），由RB组成的网络叫做TRILL Campus。

RB间互联方式

RB和RB之间可以直接连接，也可以通过一个二层交换机组成的网络互联。

TRILL概述—TRILL报文格式

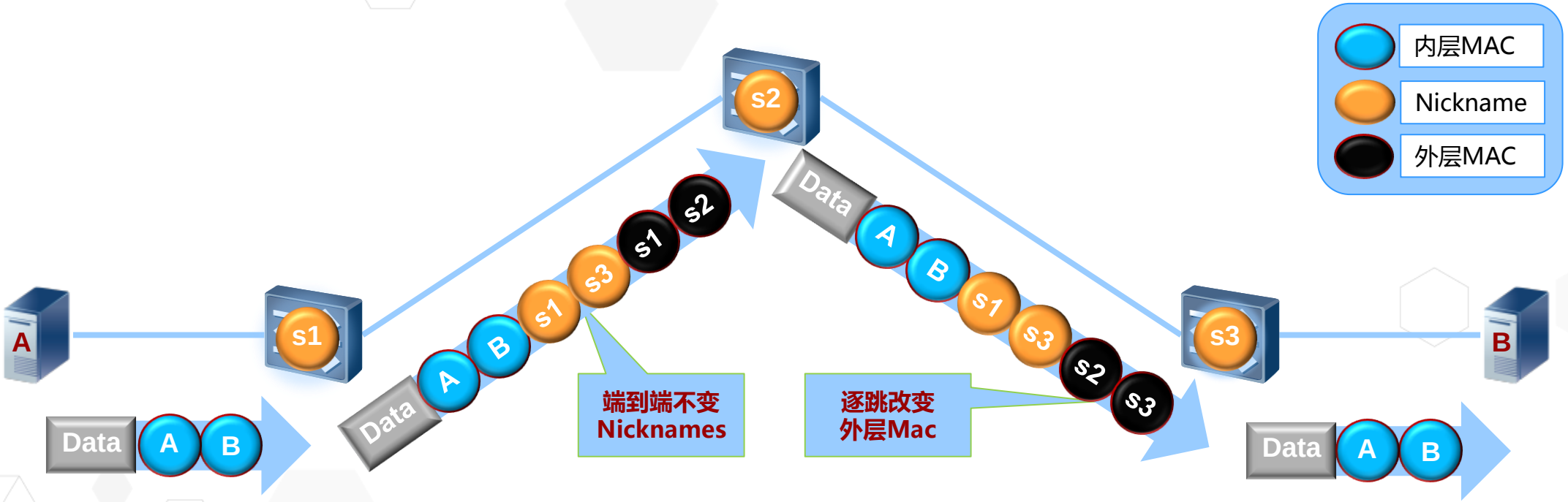


MAC – in TRILL – in MAC

TRILL报文格式

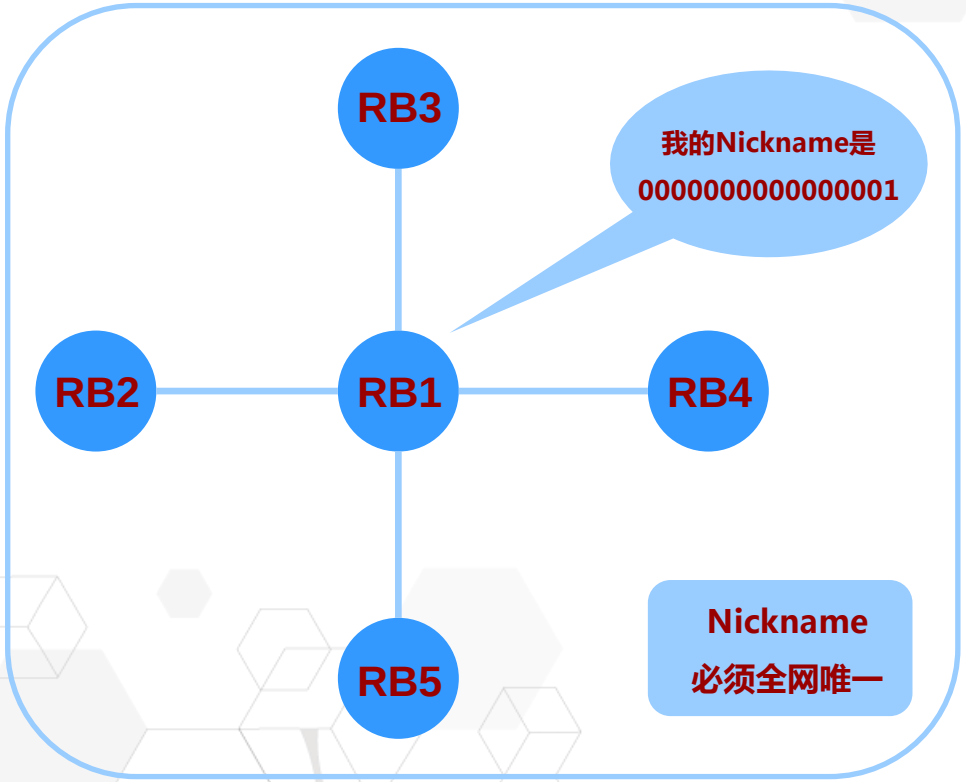
- **DA**：外层目的MAC，单播为下一跳RB的MAC，组播为保留MAC。
- **SA**：外层源MAC，为每一跳RB自身MAC。
- **VLAN**：承载TRILL数据报文的外层VLAN，为TRILL协议中指定 VLAN。
- **V**：trill版本号，当前为0，如果发现不为0的版本会直接丢弃。
- **R**：保留字段。
- **M**：组播标志，0为单播，1为组播。
- **Op-Length**：trill头扩展选项的长度。
- **Hop**：跳数。
- **E-Rb-Nickname**：单播为出口RBnickname，组播为树根nickname。
- **I-Rb-Nickname**：Ingress RB的nickname。
- **Original Frame**：服务器发出的原始二层报文。

TRILL概述— TRILL转发数据封装



源终端的原始二层报文能够穿越Trill网络到达目的终端，Trill网络对于服务器来说相当于是Bridge Fabric！

TRILL概述—Nickname概念



Nickname概念

- TRILL网络中的RB以nickname进行标识，nickname是一个2字节数值。
- RB可以有多个nickname，可自动生成，也可手工配置，须保证全网唯一。
- 跟随nickname的还有两个属性：priority（优先级）、root priority（树根优先级），分别用于nickname冲突协商和分发树树根选举。

Nickname冲突协商

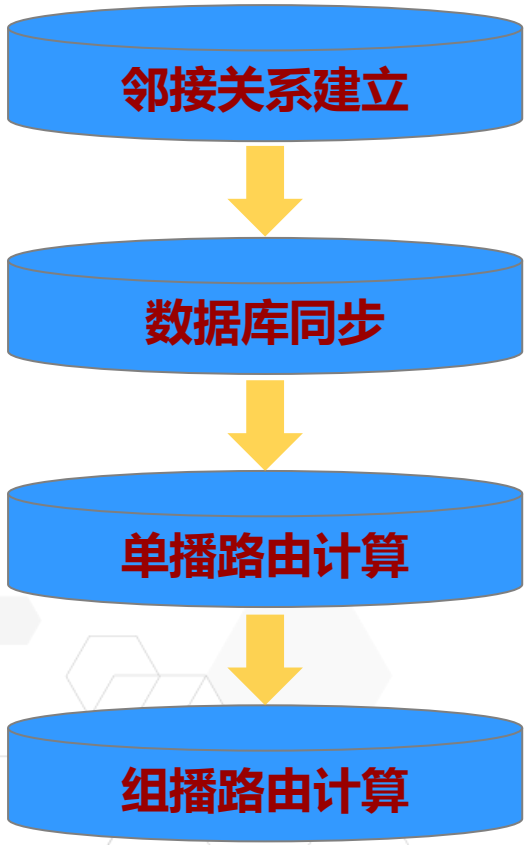
- 由于nickname可以自动生成，就可能出现两台RB产生相同的nickname，所以TRILL协议就提供了一个nickname优先级字段用于解决冲突。
- 对于新入网的RB，需等待同步完现有网络LSDB后，确认本地Nickname和现网不冲突后再发布；如果冲突，则需要重新选取，避免影响现网中已有业务。

TRILL概述—TRILL和其他几种二层技术对比

	传统二层	CSS+iStack	TRILL	SPB
封装方式	传统ETH头（无TTL）	传统ETH头（无TTL）	TRILL（有TTL）	MacInMac（无TTL）
破环方式	MSTP协议	管理方式	TRILL协议	SPB协议
ECMP	不支持ECMP	通过LAG方式，支持ECMP	和三层IP网络类似，支持逐跳ECMP	Ingress节点基于流进行ECMP，不支持逐跳ECMP
组播树数目	NA	NA	少（二层共享组播树）	多（二层源组播树）
最短路径转发	不支持	支持	支持	支持
收敛时间	长，而且收敛时间不稳定	短	较短（整网收敛几百ms）	较短（整网收敛几百ms）
多租户支持	4K（按照VLAN进行隔离）	4K（按照VLAN进行隔离）	4K（按照VLAN进行隔离），将来可以演进到通过FineLabel来隔离租户，从而可以支持16M租户	支持16M（按照I-SID进行隔离）
组网成本	低	高（框间通信带宽占用多，而且很难做到无阻塞）	低	低
网络规模	小	中等（堆叠节点数目受限，而且堆叠框架无法做到无阻塞）	大	大
适合组网	适合逐级收敛的组网，不适合扁平化胖树组网	适合扁平化胖树组网	适合扁平化胖树组网	比较适合扁平化胖树组网，更适合于IPTV里面点到多点组网方式

目录

TRILL协议整体过程



邻接关系建立

■邻居发现、握手、最后邻居状态up。

■广播链路上，还需要选举DRB、进行端口角色通告、指定AF、指定DesignatedVLAN

数据库同步

■所有设备都拥有整网所有设备系统ID、Nickname及其属性、Ingress RB的interestedVLAN（接入VLAN）、Neighbour TLV。

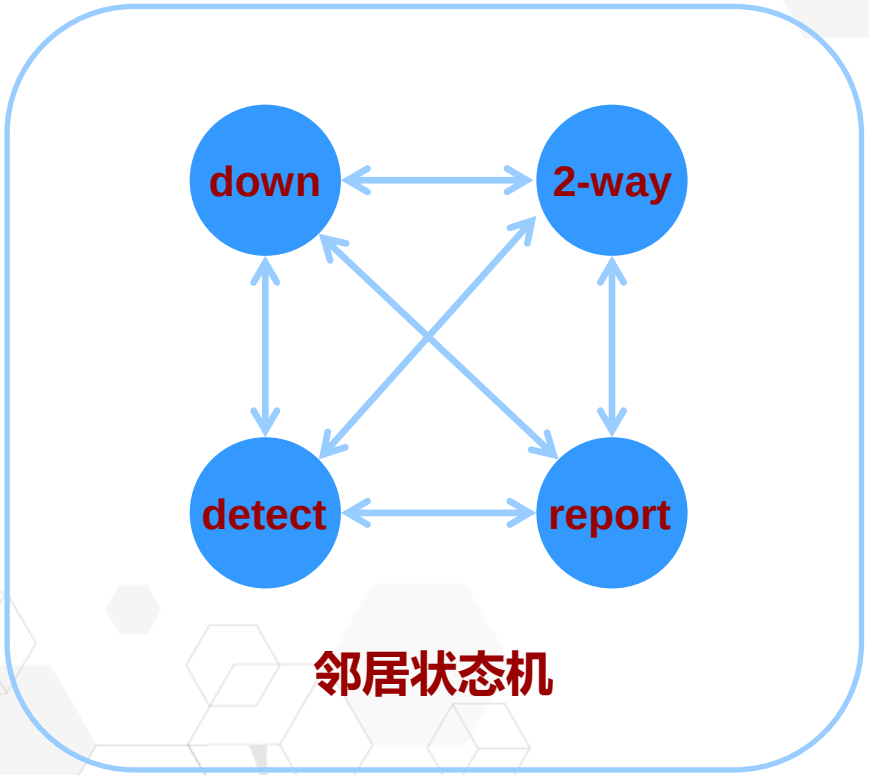
单播路由计算

■每台设备以自身作为源节点，计算到达所有其他节点的最短路径树。

组播路由计算

■设备以分发树树根为源节点，计算到达其他节点的最短路径树，再根据各个Ingress RB通告的接入VLAN信息，执行剪枝计算，生成剪枝后的分发树表项。

一：TRILL邻接关系管理—邻居状态协商



邻居状态机

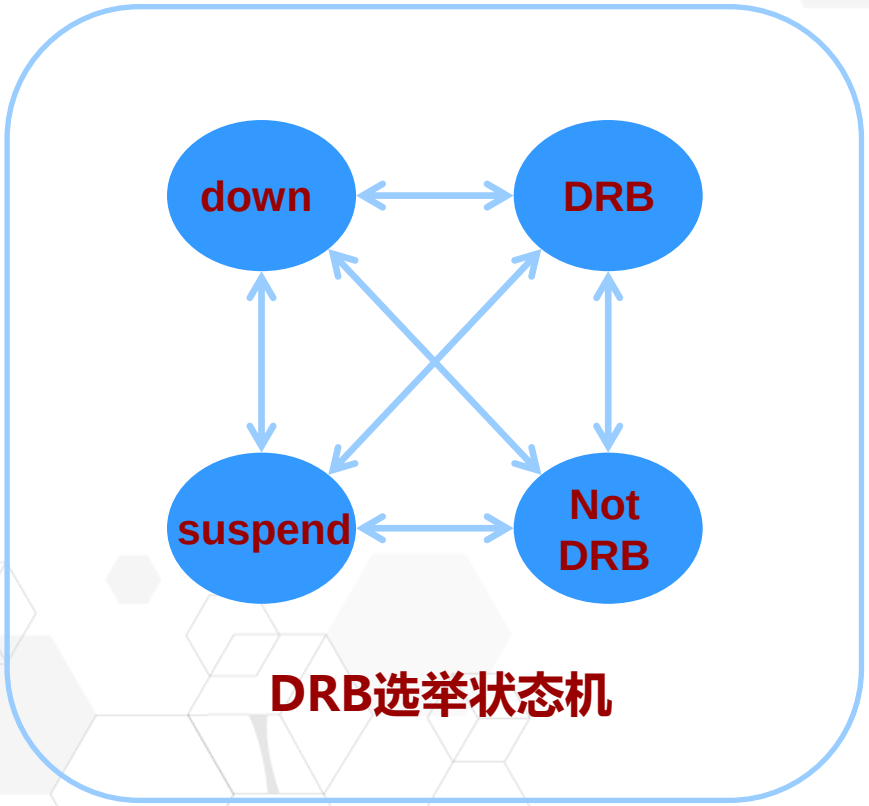
Hello报文

- Hello报文用于广播链路上TRILL邻接关系协商、 DesignatedVLAN指定、DRB选举、端口角色通告、MTU探测等，运行在多归接入端口上还支持AF功能。

邻居状态

- Down：初始状态，表明邻居不存在。
- Detect：检测到邻居存在，但是还没有握手成功。
- 2-WAY：握手成功，如果使能MTU探测，MTU探测还没有完成。
- report：握手成功，MTU探测完成。

一： TRILL邻接关系管理—DRB选举



DRB选举状态机

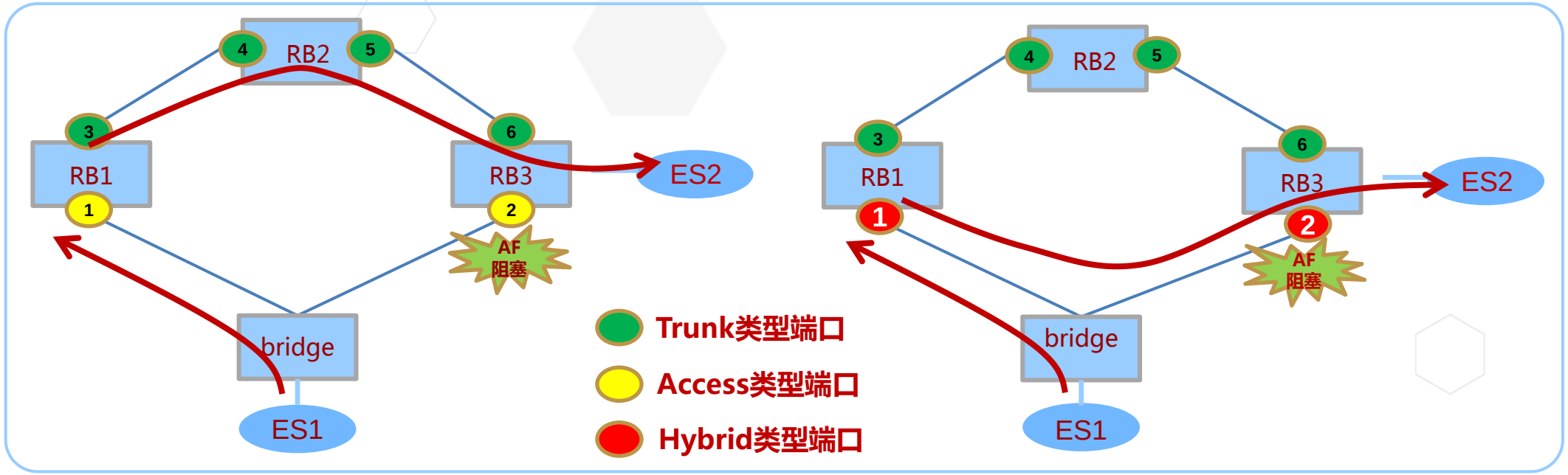
DRB功能

- 指定本Link上用于传送TRILL数据报文的Designated-VLAN 。
- 在接入端口还可以进行VLAN Forwarder指定。
- 决定是否创建Pseudonodes，Pseudonodes在Link上有多个RB连接时，可以将RB之间FullMesh的连接关系变为星型连接关系，减少通告的LSP数量。若Link上只有两个RB，则DRB可以在 Hello 报文中设置bypass pseudonode bit，表明不创建Pseudonodes。

邻居状态

- Down：端口链路状态down或没有使能TRILL 。
- suspend：收到跟本端MAC地址相同的TRILL-Hello报文，但是自身DRB优先级较低，该状态和Down状态类似。
- DRB： DRB端口，可以收发TRILL数据报文。
- NotDRB：非DRB端口，可以收发TRILL数据报文。

一： TRILL邻接关系管理—端口角色

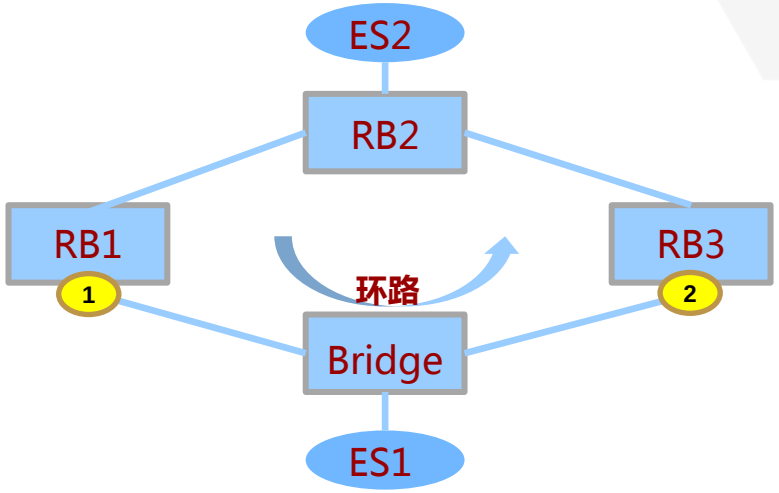


三种端口角色

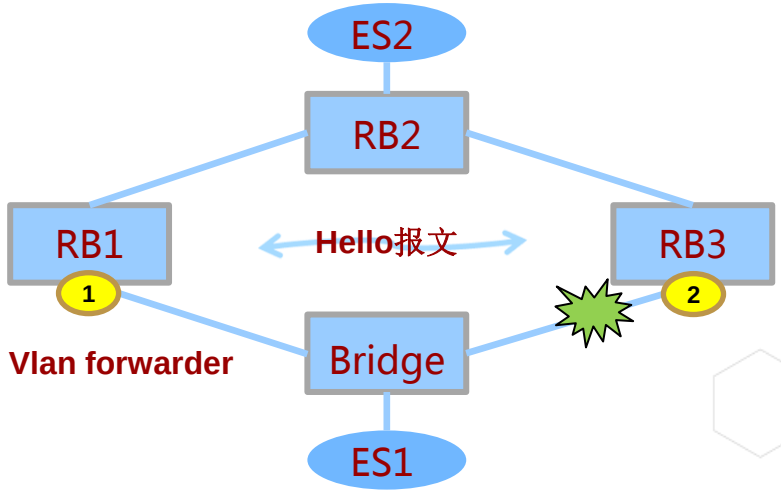
- Access：UNI接口。该类型端口用于接入用户终端，只能转发Native Ethernet数据报文，不能转发TRILL数据报文，链路不会通过LSP协议报文发布出去。只有该类型端口需要进行AF选举。
- Trunk：NNI接口。支持广播链路。只允许转发TRILL数据报文和协议报文，不允许转发Native Ethernet数据报文。
- P2P:NNI接口。相比Trunk类型端口，不进行DRB选举，其他和Trunk类型端口作用一样。

注：端口使能Trill协议后，缺省为Hybrid端口，即既是Access又是Trunk端口，该端口既能接入终端的Native ETH报文，又能转发TRILL数据报文。

一： TRILL邻接关系管理—AF选举



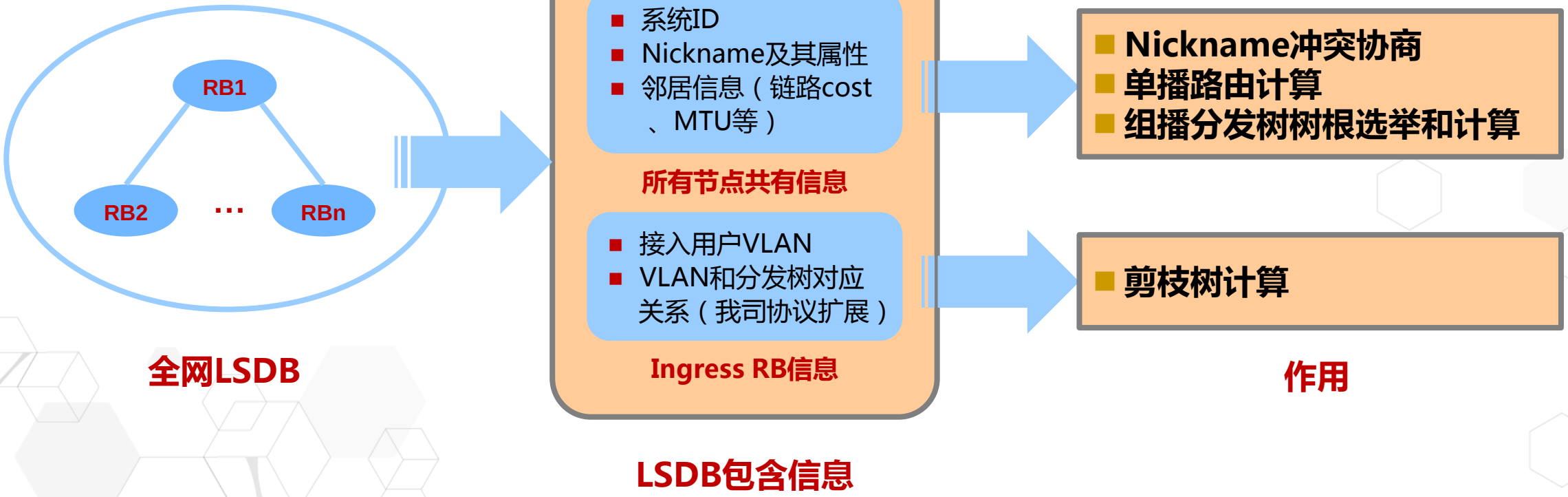
ES1经传统二层交换机双归RB



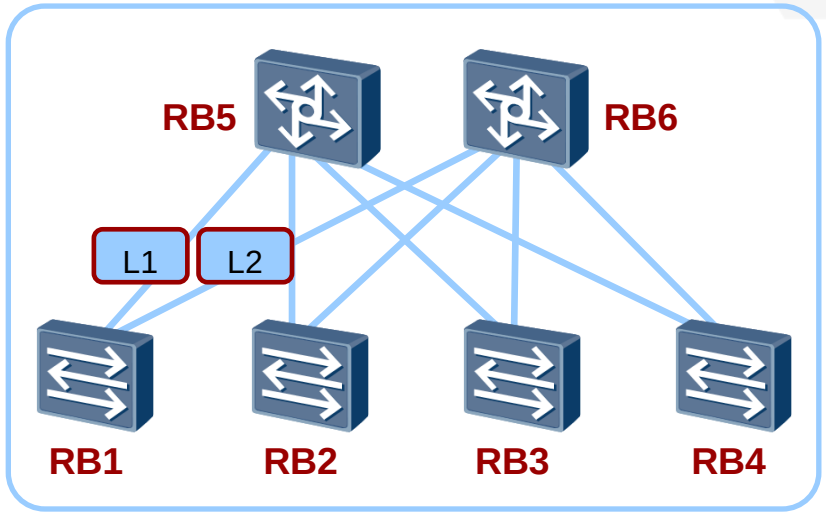
ES1经传统二层交换机双归RB

无AF功能模块	引入AF功能模块
VLAN内未知单播或广播报文会通过TRILL网络形成环路，会形成环路风暴。并且由于在接入侧端口已经剥去TRILL封装，没有TRILL头部HopCount进行保护，对网络的破坏更大。	通过在TRILL接入侧端口（Access Port）之间运行TRILL Hello协议，由DRB指定某一台RB（RB1）作为接入用户的VLAN Forwarder，其他都是非VLAN Forwarder，这样就不会在接入侧形成二层环路了。

二：TRILL全网数据库同步



三：TRILL路由收敛—单播路由表生成



- 所有链路cost值相同
- RB1到RB6的系统MAC分别为MAC1-6
- RB1到RB6的nickname分别为Nickname1-6

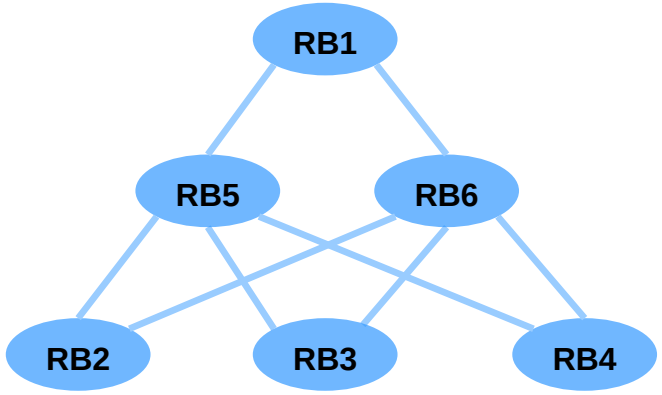
① SPT计算

- 通过全网LSDB，以本设备为源节点，生成到所有其他节点的SPT树。

② 生成nickname单播路由

- 结合邻居信息，获取到达邻居节点的出接口、下一跳。
- 根据每个节点发布的nickname，生成nickname单播转发表。

三： TRILL路由收敛—RB1上单播Nickname路由表生成



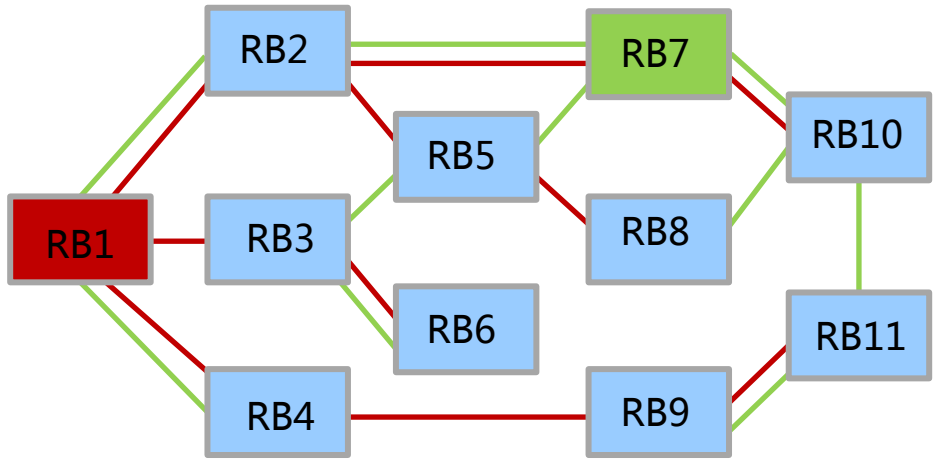
RB1到其他节点最短路径树

RB1邻居表项	
邻居	出接口、下一跳MAC
RB5	L1、MAC5
RB6	L2、MAC6

RB1设备单播nickname转发表	
目的Nickname	出接口和下一跳MAC
nickname2	L1、MAC5 L2、MAC6
nickname3	L1、MAC5 L2、MAC6
nickname4	L1、MAC5 L2、MAC6
nickname5	L1、MAC5
nickname6	L2、MAC6

- RB1生成到所有其他节点的最短路径树，保证了单播流量转发为最短路径。
- RB1到RB2、RB3、RB4都有两条等价最短路径，因此形成了单播流量负载分担，链路带宽利用率更高。
- 进来的数据报文基于TRILL头部的EgressNickname查找单播nickname转发表，获取出接口和下一跳信息，如果存在多个出接口，则根据ECMP算法选择一个出接口，转发出去。

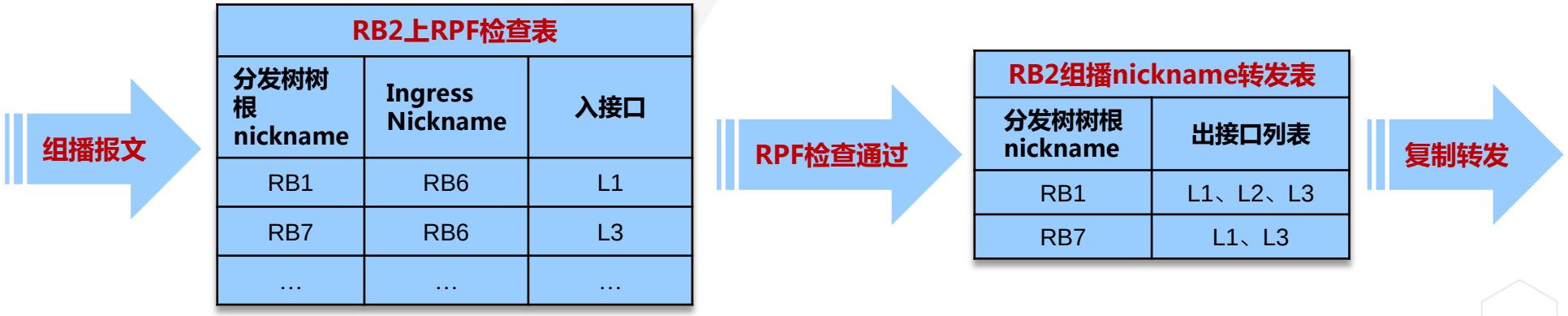
四：组播路由计算



- **Root priority最高设备选举**：每台设备根据整网所有设备发布的Nickname的root priority和支持的分发树数目，获取root priority最高的Nickname以及整网最小的分发树数目N 。
- **分发树树根选择**：拥有root priority最高的Nickname所在RB有权指定哪些nickname是分发树树根，如果没有指定则以root priority最高的N个Nickname为分发树树根。
- **分发树计算**：分别以N个分发树树根为源节点，计算到整网所有其他节点的最短路径树。
- **RPF检查表生成**：基于每个Ingress RB通告的选择的分发树信息，生成RPF检查表，用于避免组播环路。
- **剪枝计算**：基于每个Ingress RB通告的接入VLAN信息，进行剪枝计算，节省TRILL网络内带宽。

注： root priority最高的节点、分发树树根必须单播可达，因此组播路由计算需要在单播计算之后。

四：组播路由转发表和RPF检查表



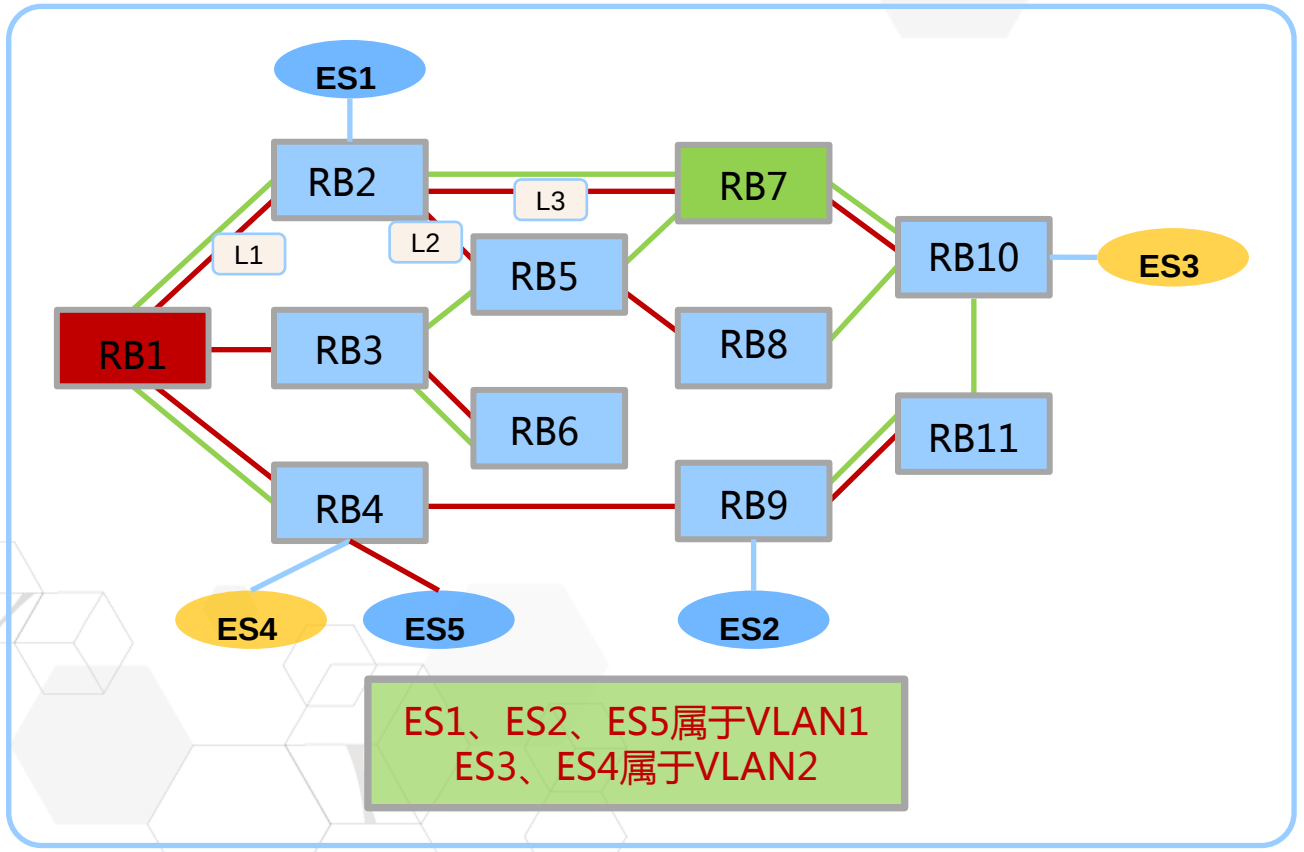
RPF检查

- Ingress设备RB6选择分发树1转发组播数据流量，到达RB2之后，该报文只能从L1进入，如果从其他端口进来，RPF检查失败，报文丢弃。
- 由于TRILL网络采用共享树，所有Ingress RB都可能为组播流量入口RB，因此RB需要基于分发树树根Nickname和Ingress RB的Nickname、入端口生成RPF检查表项，根据该表项对相应的组播流量进行RPF检查，防止TRILL网络内部组播报文转发出现环路。

组播路由转发

- 组播数据报文转发时候芯片需要做源端口剪枝，比如从L1端口进来的报文不能再从L1复制出去，只能向出接口列表的剩余接口转发。
- 每台RB都基于这些树根独立计算组播分发树，由于都拥有统一的全网拓扑，采用相同的算法，因此所有RB对于组播树的形态理解都是一致的，这样即使没有PIM网络中的Join、Prune消息，也能够建立统一的分发树表项。

四：基于VLAN的分发树剪枝计算



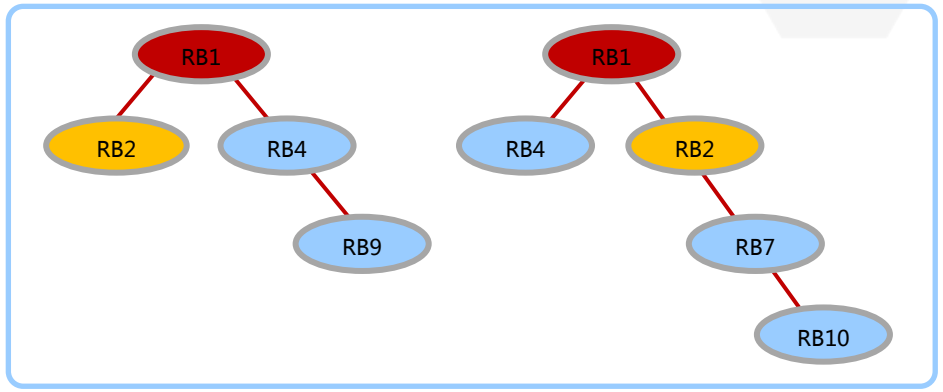
Ingress RB发布信息

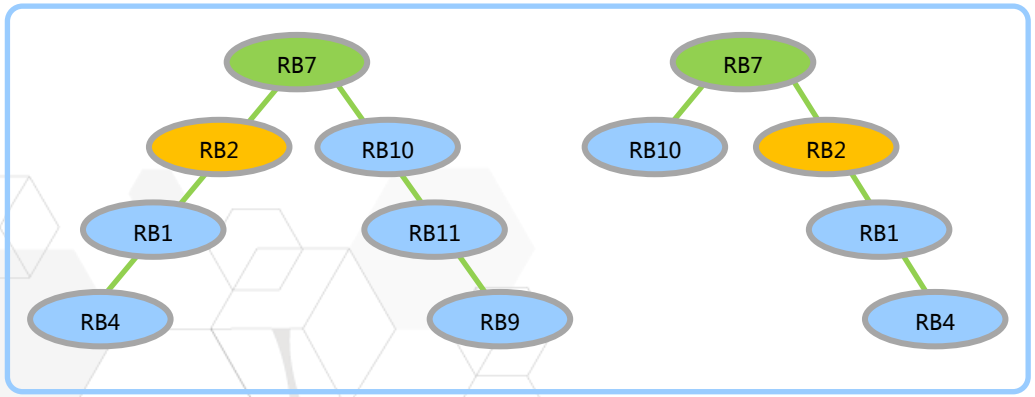
- 左图中RB2、RB4、RB9发布接入VLAN1，RB4、RB10会发布接入VLAN2。

所有其他RB基于VLAN进行剪枝计算

- 所有其他RB基于每个Ingress RB发布的信息进行剪枝计算，生成剪枝后的分发树表项；TRILL网络内组播流量只向相应的边缘RB复制，实现按需转发，节省TRILL网络带宽。

四：基于VLAN的分发树剪枝计算



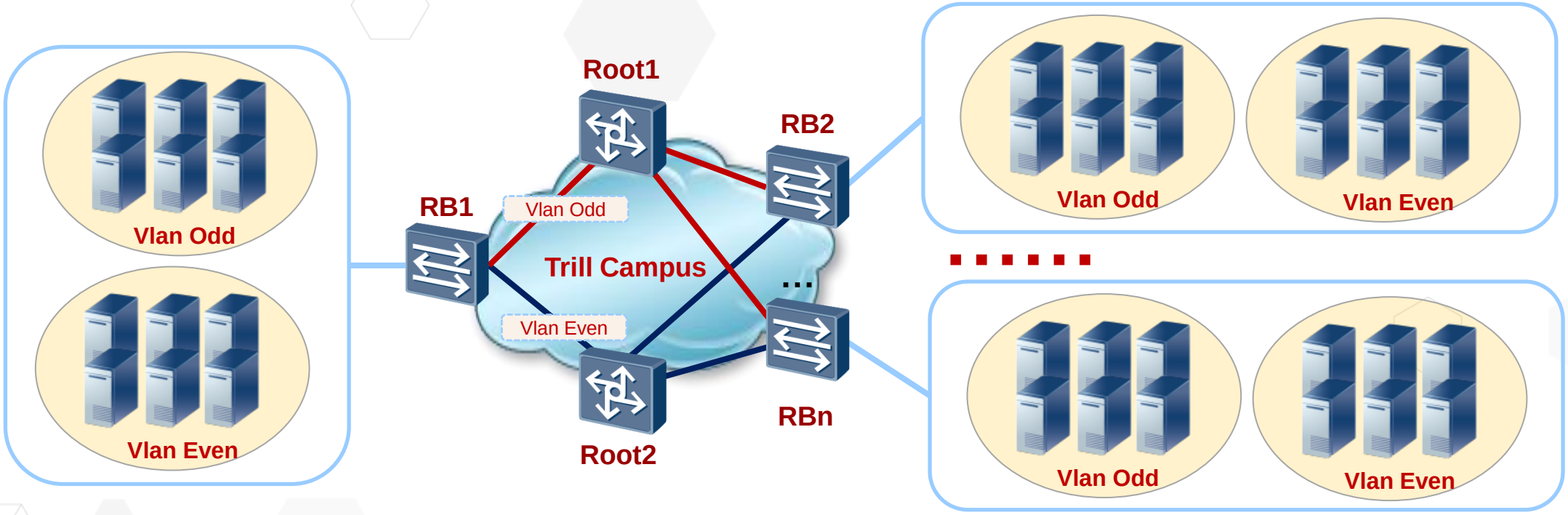


RB2设备上组播剪枝表项		
分发树树根nickname	接入VLAN	出接口
RB1	VLAN1	L1
RB1	VLAN2	L1、L3
RB7	VLAN1	L1、L3
RB7	VLAN2	L1、L3

组播路由转发

- 组播数据报文转发时候基于TRILL头部目的Nickname+内层VLAN作为key，查找基于VLAN进行剪枝的组播表项。
- RPF检查不受组播剪枝计算影响，仍然是基于树根Nickname+Ingress RB的Nickname进行检查。

四：组播流量负载均衡



◆Ingress RB为不同接入VLAN选择不同的分发树，比如上图，对于奇数内组播流量选择Root1所在分发树进行转发，对于偶数VLAN内组播流量选择Root2所在分发树进行转发，使整网组播流量能够基于接入VLAN实现负载均衡。

四：分发树剪枝表项数目优化

普通方法生成的分发树剪枝表项		
分发树树根RB	VLAN ID	出接口
Root1	1	L1
…	…	…
Root1	1000	L1
Root1	1001	L1
…	…	…
Root1	2000	L1
Root2	1	L2
…	…	…
Root2	1000	L2
Root2	1001	L2
…	…	…
Root2	2000	L2

优化的分发树剪枝表项

我们方法生成的分发树剪枝表项		
分发树树根RB	VLAN ID	出接口
Root1	1	L1
…	…	…
Root1	1000	L1
Root2	1001	L2
…	…	…
Root2	2000	L2

分发树剪枝表项规模和分发树数目无关！

Ingress RB发布信息

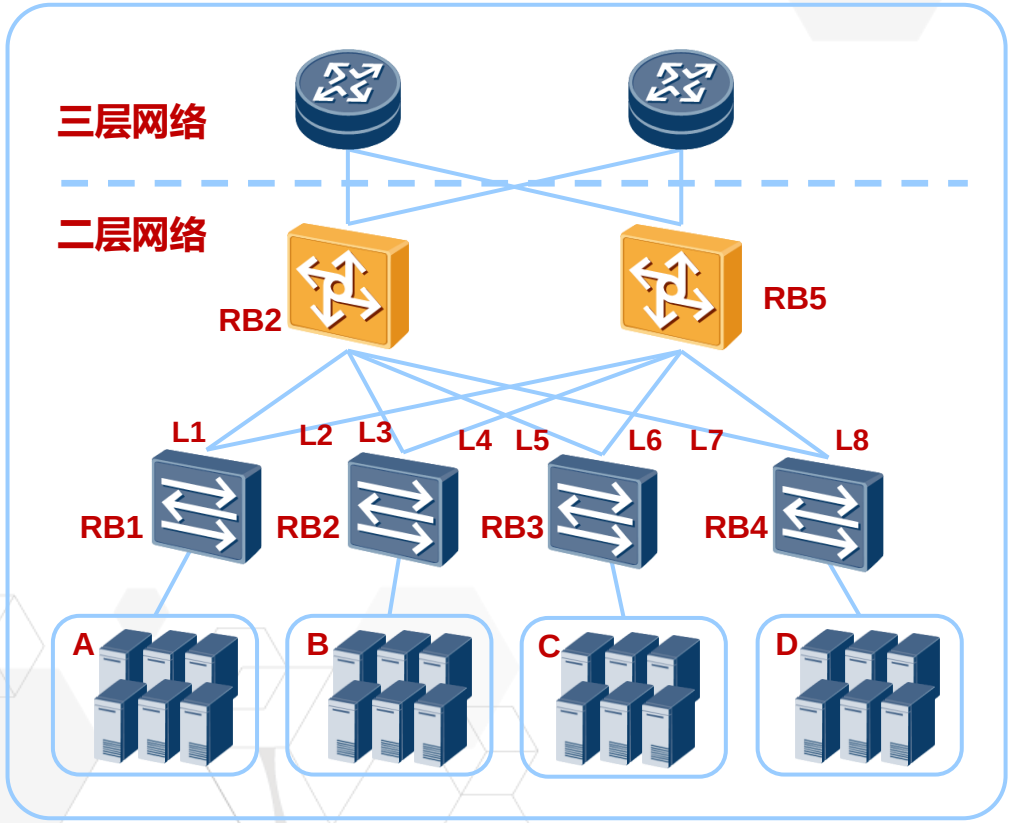
- Ingress RB基于VLAN进行组播流量负载均衡处理
- Ingress RB发布VLAN和树根对应关系，RB1发布VLAN 1-1000和Root1对应关系，RB2发布VLAN 1001-2000和Root2对应关系。

其他RB基于VLAN进行剪枝计算

- 所有其他RB基于每个Ingress RB发布的VLAN、对应的分发树进行剪枝计算，而不是基于每个分发树进行剪枝计算。

目录

TRILL数据转发总体流程

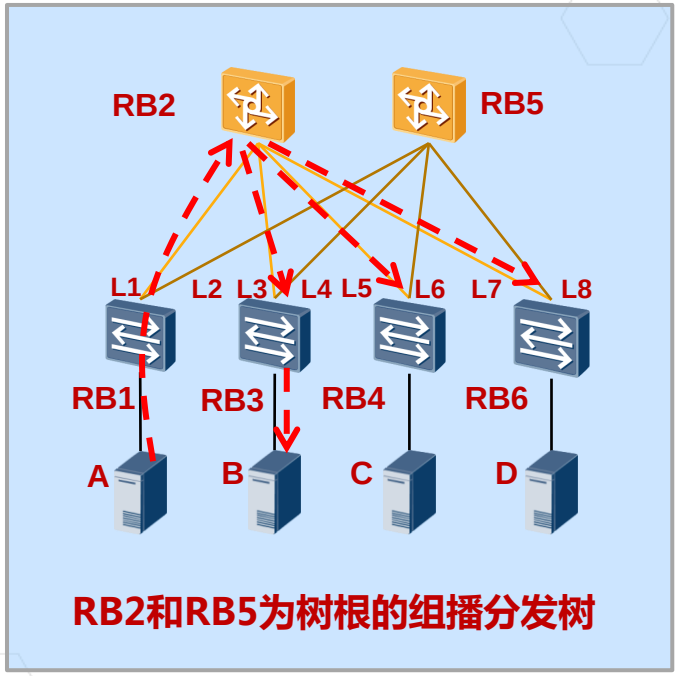


总体流程

- HOST A发送ARP Request报文到HOST B
- HOST B回应ARP Reply报文到HOST A。
- HOST A发送单播数据报文到HOST B。

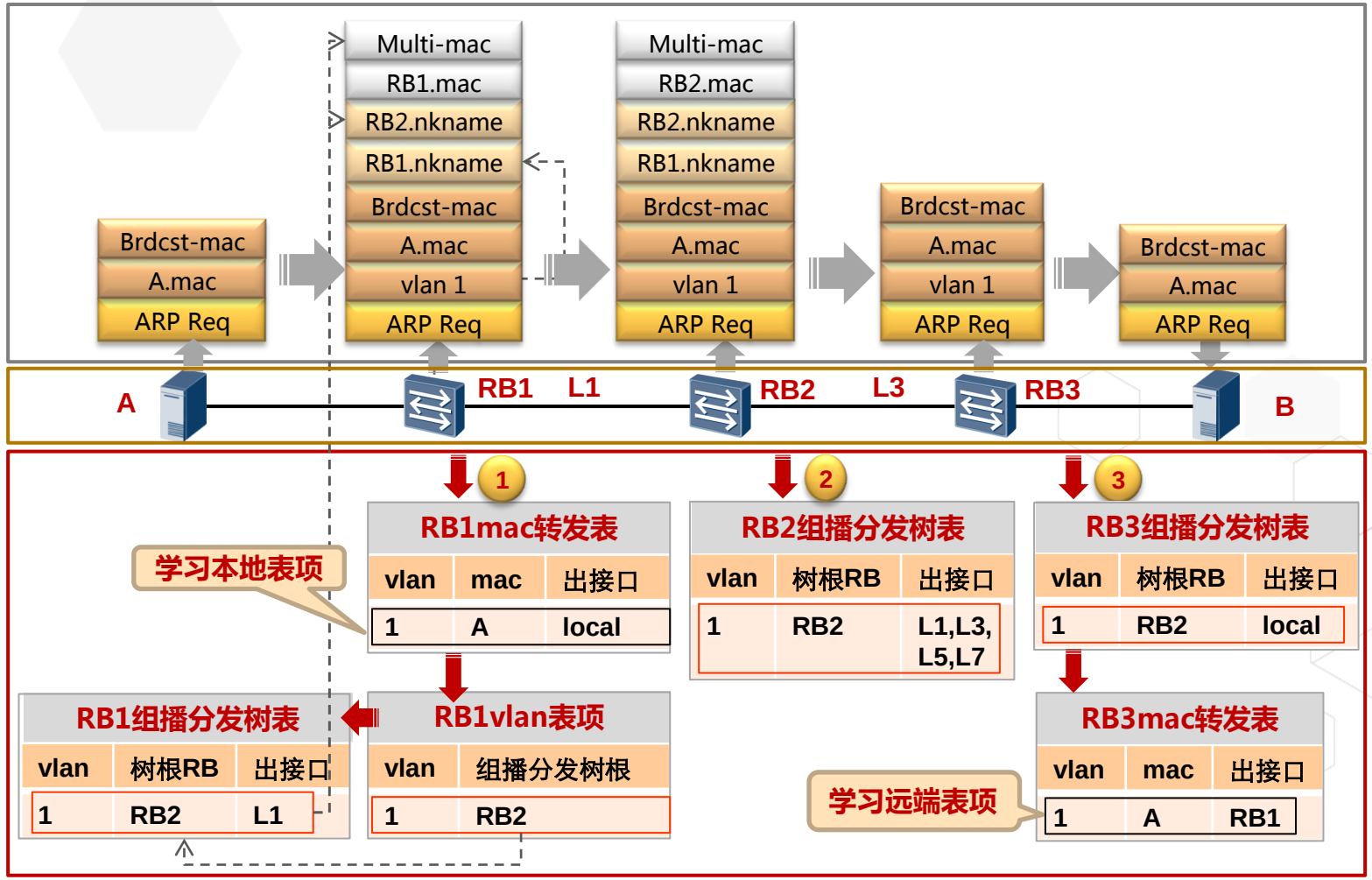
注：A、B、C、D都属于VLAN 1。对于VLAN内组播或广播报文转发流程，和ARP Request报文转发流程一样。

ARP Request转发流程

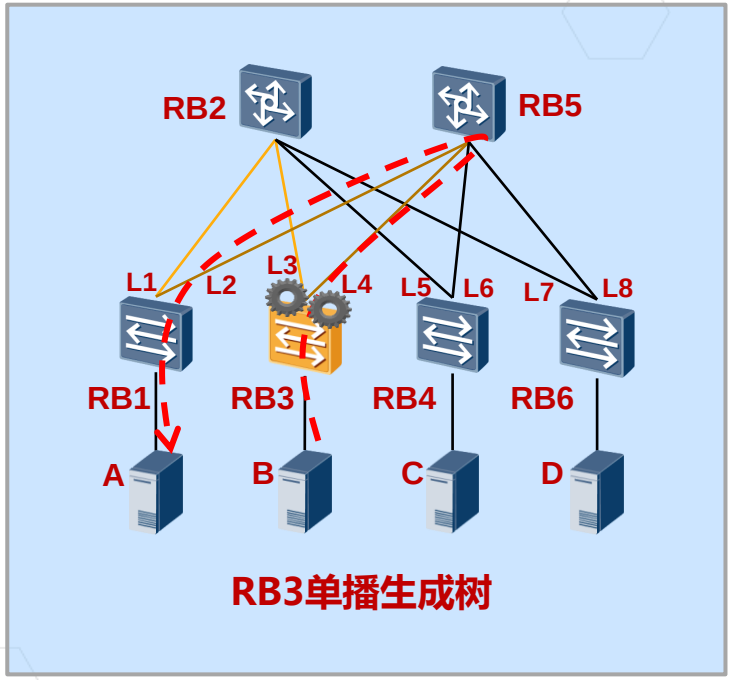


RB2和RB5为树根的组播分发树

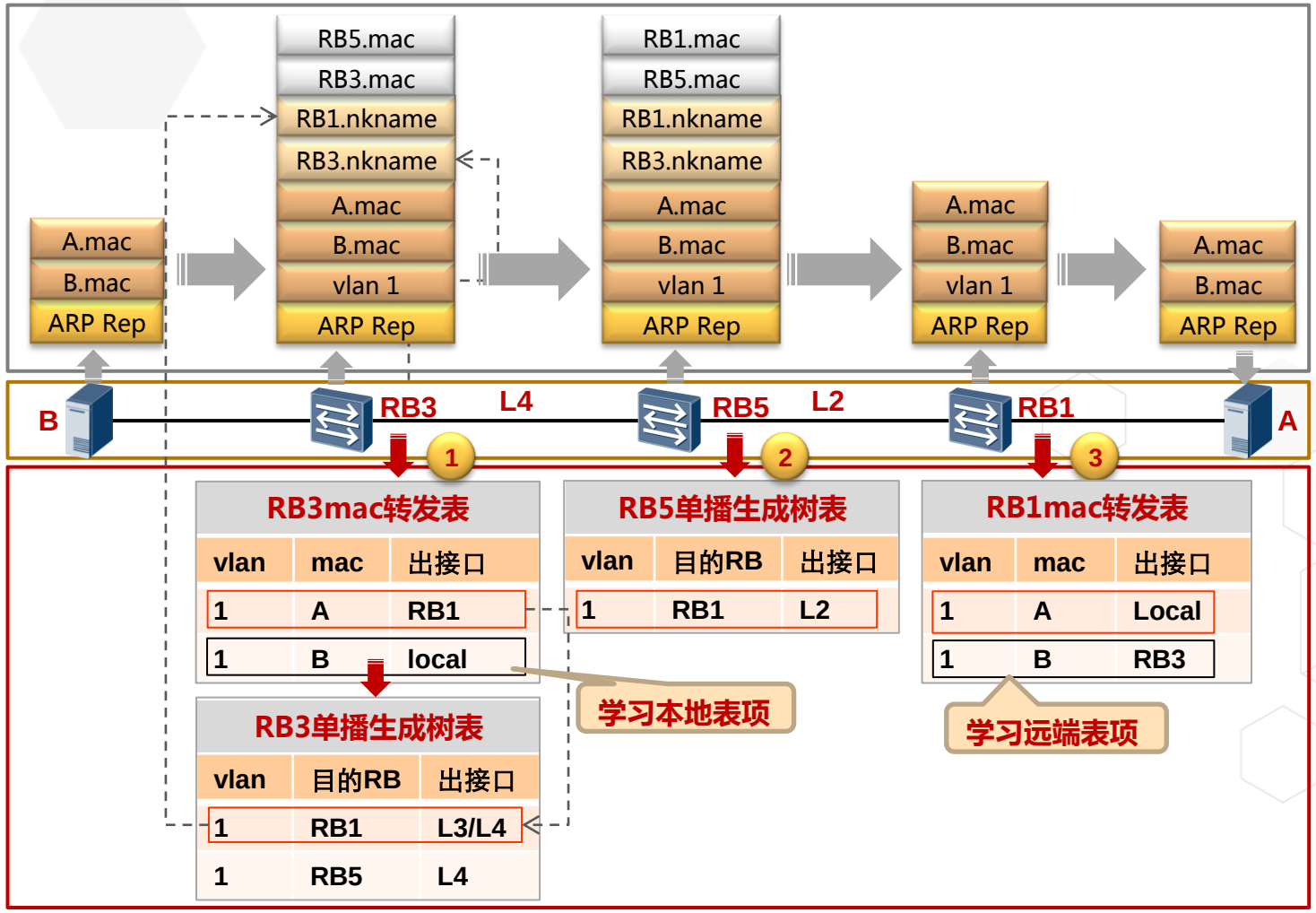
1. Ingress RB查不到mac转发表项，根据VLAN对应的分发树树根查找组播转发表，封装并转发ARP Request报文
2. 树根RB向所有组播成员RB发送报文
3. Egress RB接收报文解封装后本地广播



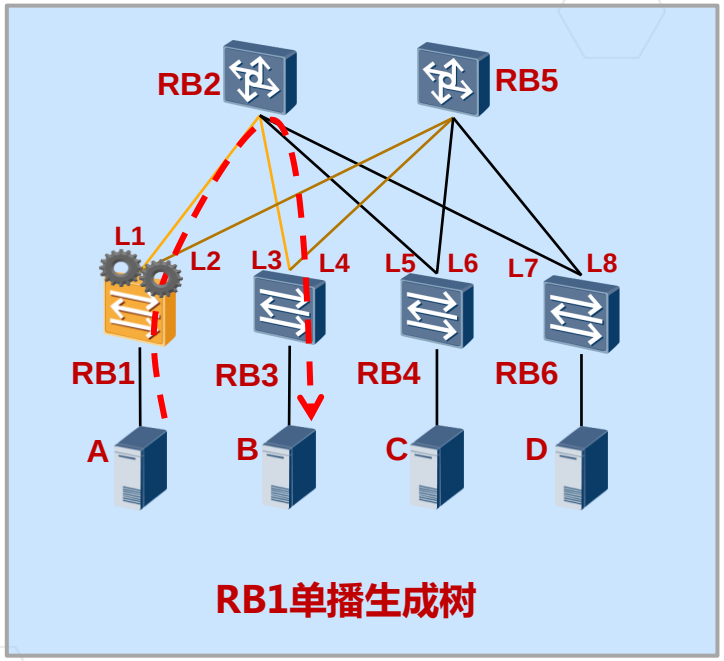
ARP Reply转发流程

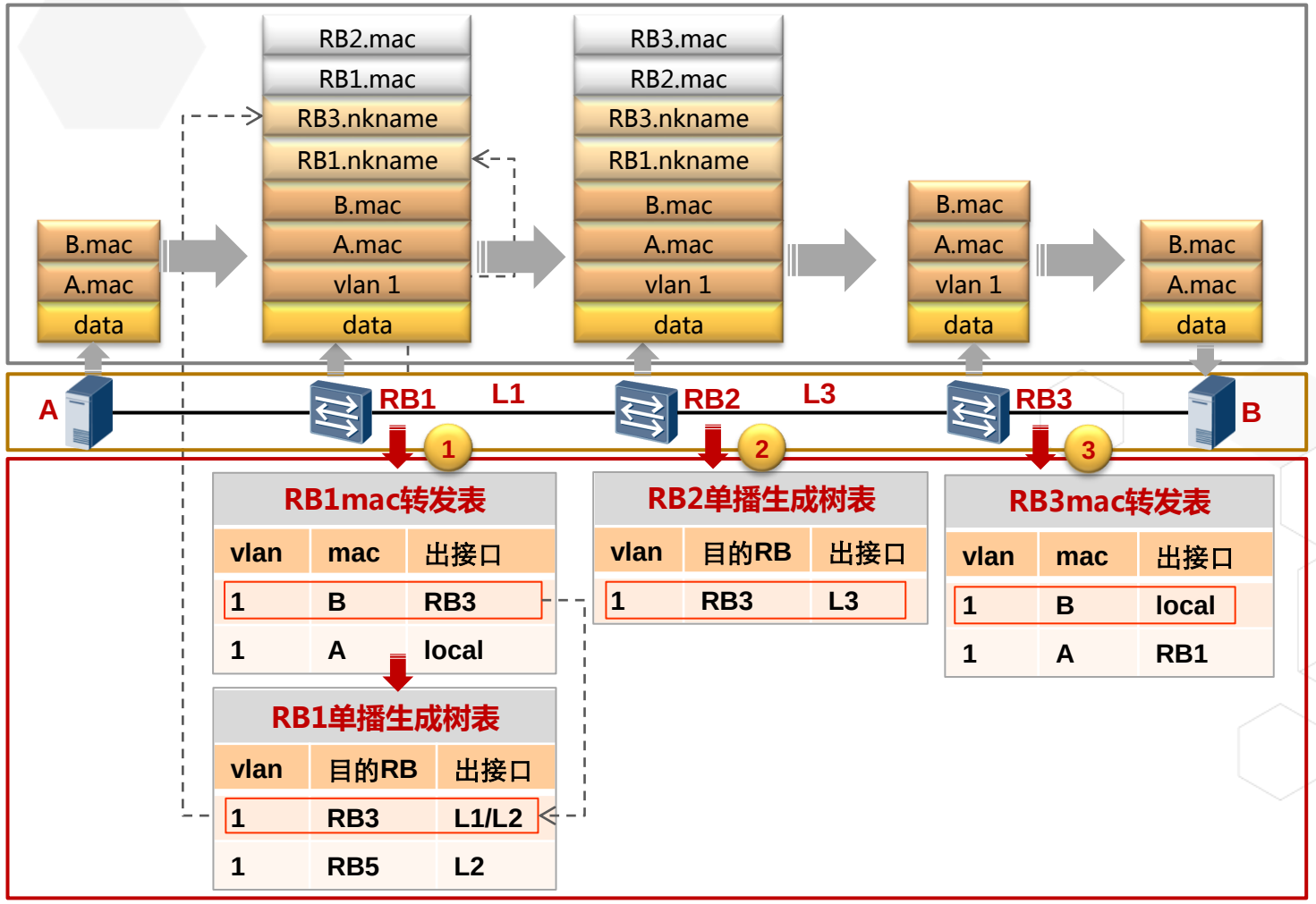


1. Ingress RB查找mac转发表，根据目的RB选择一条负载分担链路，封装转发ARP报文
2. 沿途RB根据单播生成树表项转发报文
3. Egress RB接收报文并解封装，在本地转发



A到B单播转发流程

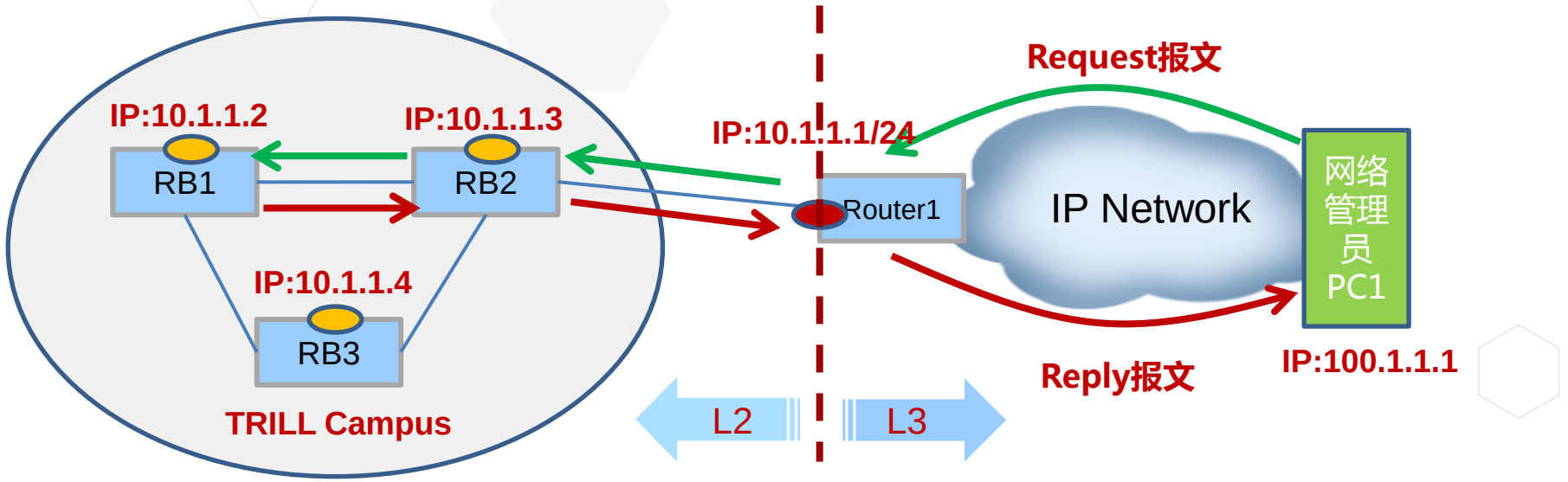




1. Ingress RB查找mac转发表，根据目的RB选择一条负载分担链路，封装转发单播报文
2. 沿途RB根据单播生成树表项转发报文
3. Egress RB接收报文并解封装，在本地转发

目录

TRILL网络设备带内网络管理

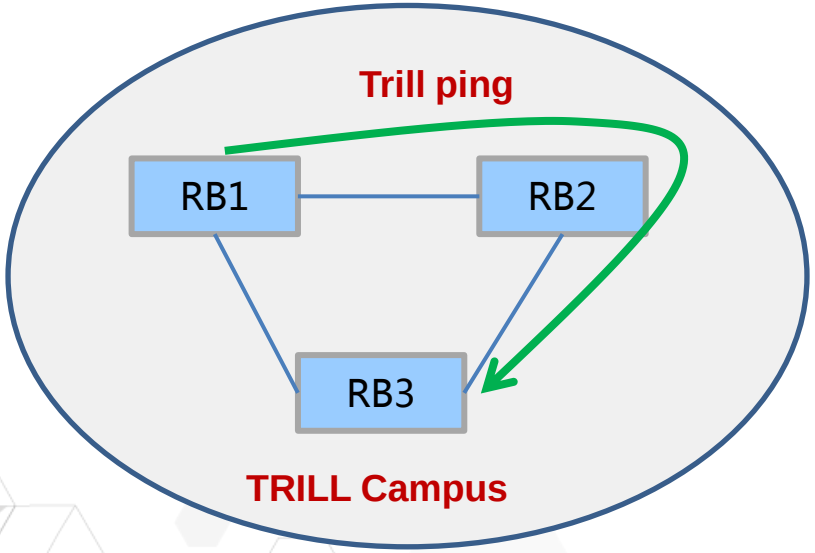


管理VLAN的VLANIF接口，管理VLAN为TRILL封装的内层VLAN

路由器对应的管理VLAN子接口

- 各RB具有内层管理VLAN对应的VLANIF接口。
- 路由器将管理VLAN子接口对应的网段10.1.1.0/24发布出去，网络管理员通过三层IP网络到达TRILL网络的出口路由器Router1，然后通过Trill网络登陆RB设备的带内管理VLANIF接口。
- 支持Telnet、SNMP、NetConf类型，通过带内网络访问RB设备进行配置管理。

TRILL网络设备故障定位



◆在RB设备上通过trill ping来检测TRILL网络转发路径连通情况，协议报文承载在专门的Trill OAM Channel之内。

Ping报文流程

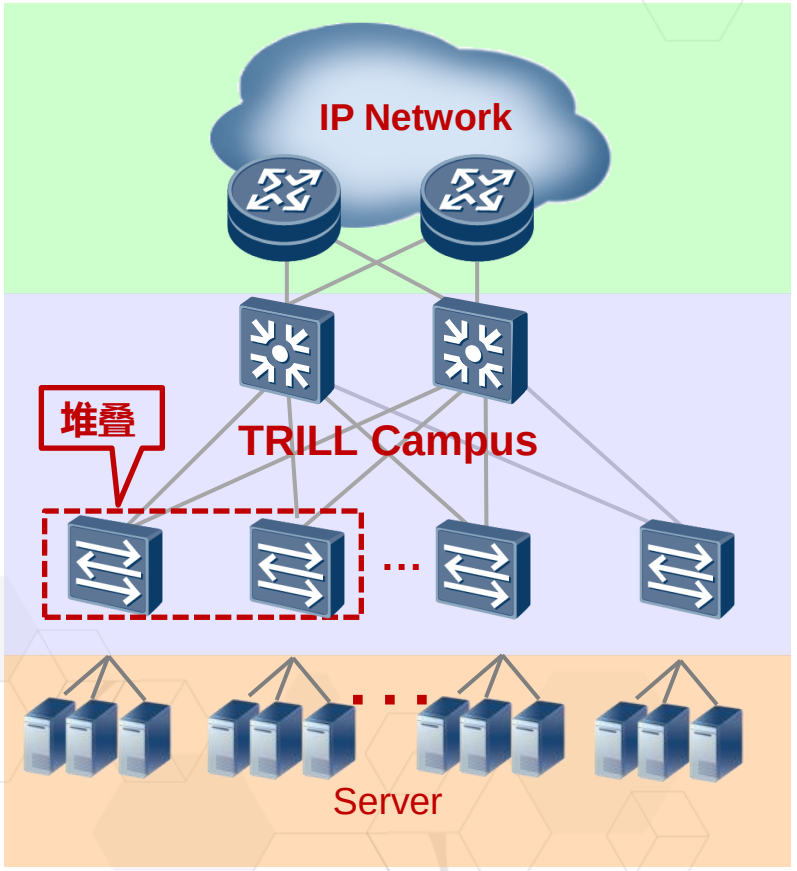
①在发起端RB上，用户指定目的nickname、超时时间、Hop-Count，发起ping，报文通过查找单播nickname转发表发送出去；

②中间节点查找nickname转发表，直到TTL=1或到达目的节点，在这些节点将报文上送CPU处理平面。

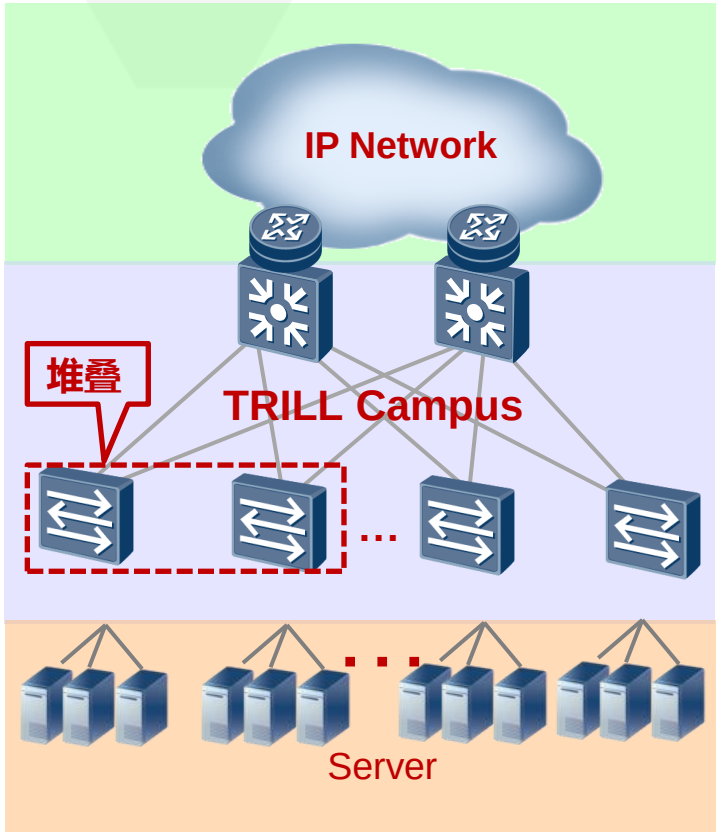
③CPU处理平面发现TTL=1，判断自身是否为目的nickname，如果为自身则回应Echo Reply；否则，回应Error Notification，原因值为TTL超时。

目录

TRILL组网应用—网关部署和服务器接入方式



组网一：三层网关和汇聚交换机分离



组网二：三层网关和汇聚交换机合一

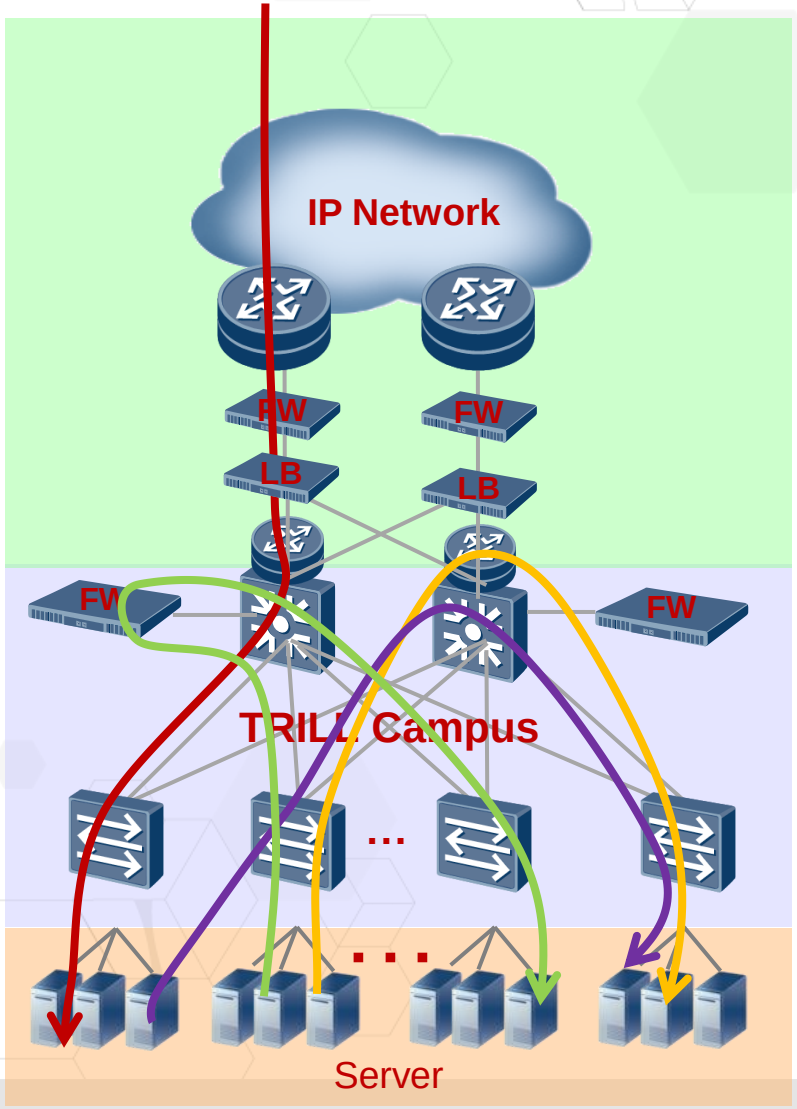
网关部署方式

- 核心RB和三层网关分开部署。
- 核心RB和三层网关合一部署：通过虚拟化技术VS，将整台设备划分为两个VS，一个VS实现三层网关功能，一个VS实现RB功能。

TRILL网络部署方式

- 接入交换机可以为TOR或者EOR，TRILL网络能够部署到直接接入服务器的接入交换机，覆盖整个DC 。
- 接入交换机支持堆叠，服务器可以通过双归方式接入，增强业务可靠性。

TRILL组网应用—增值业务部署



Client到server南北方向数据流量

跨网段Server之间流量，需要经过防火墙过滤

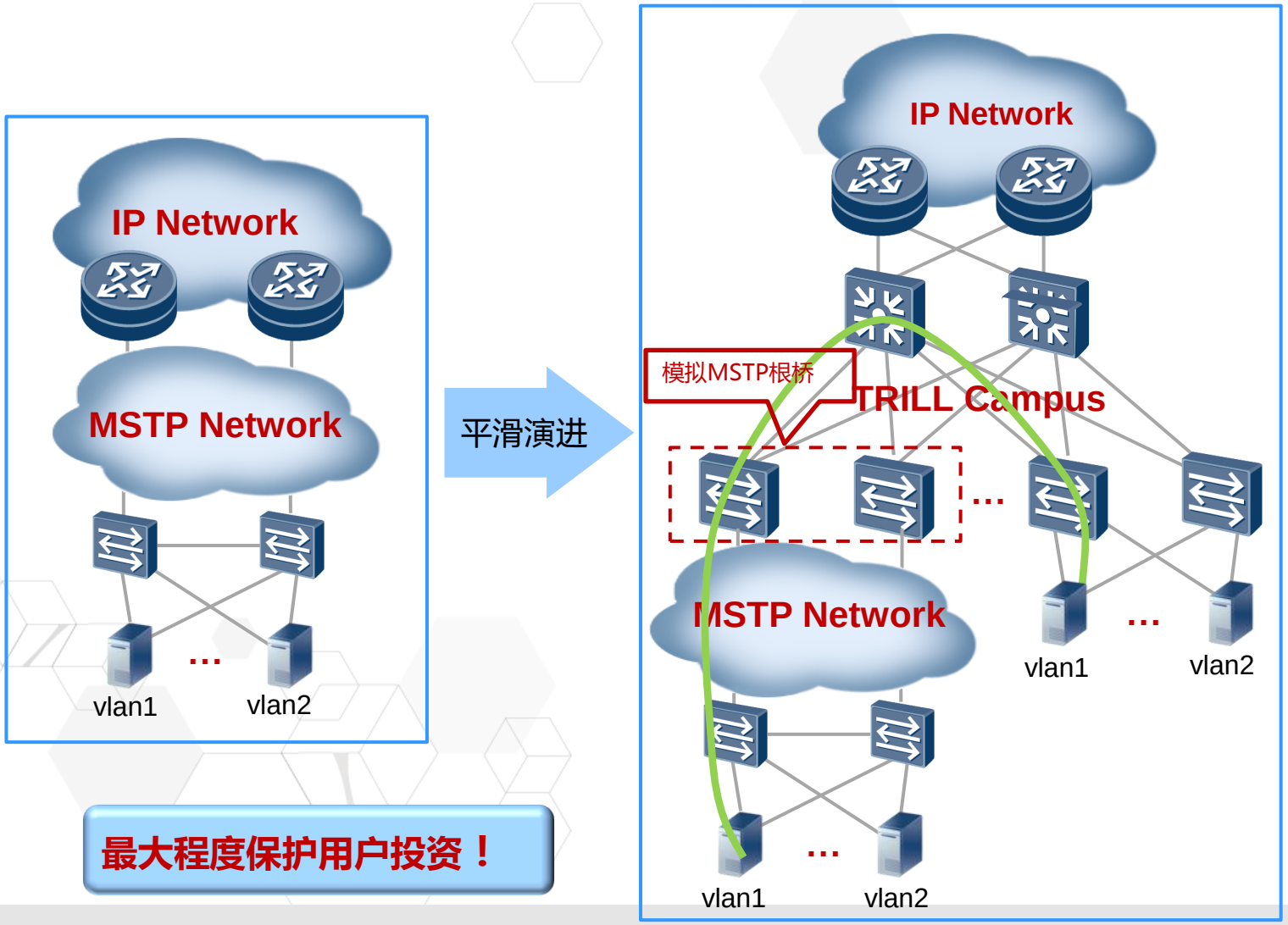
跨网段server之间流量，不需要经过防火墙过滤

同网段server之间流量

部署特点

- 对于南北向流量，FW和LB串接在汇聚交换机和出口路由器之间。
- 对于东西向跨网段流量，有些服务器属于不同安全域，需要经过防火墙进行过滤，这种情况可以将网关设置在FW上。如果属于相同安全域，则网关在汇聚交换机上。
- 对于不跨网段的东西向流量，直接经过TRILL网络进行二层转发。

TRILL组网应用—DC平滑演进一



最大程度保护用户投资！

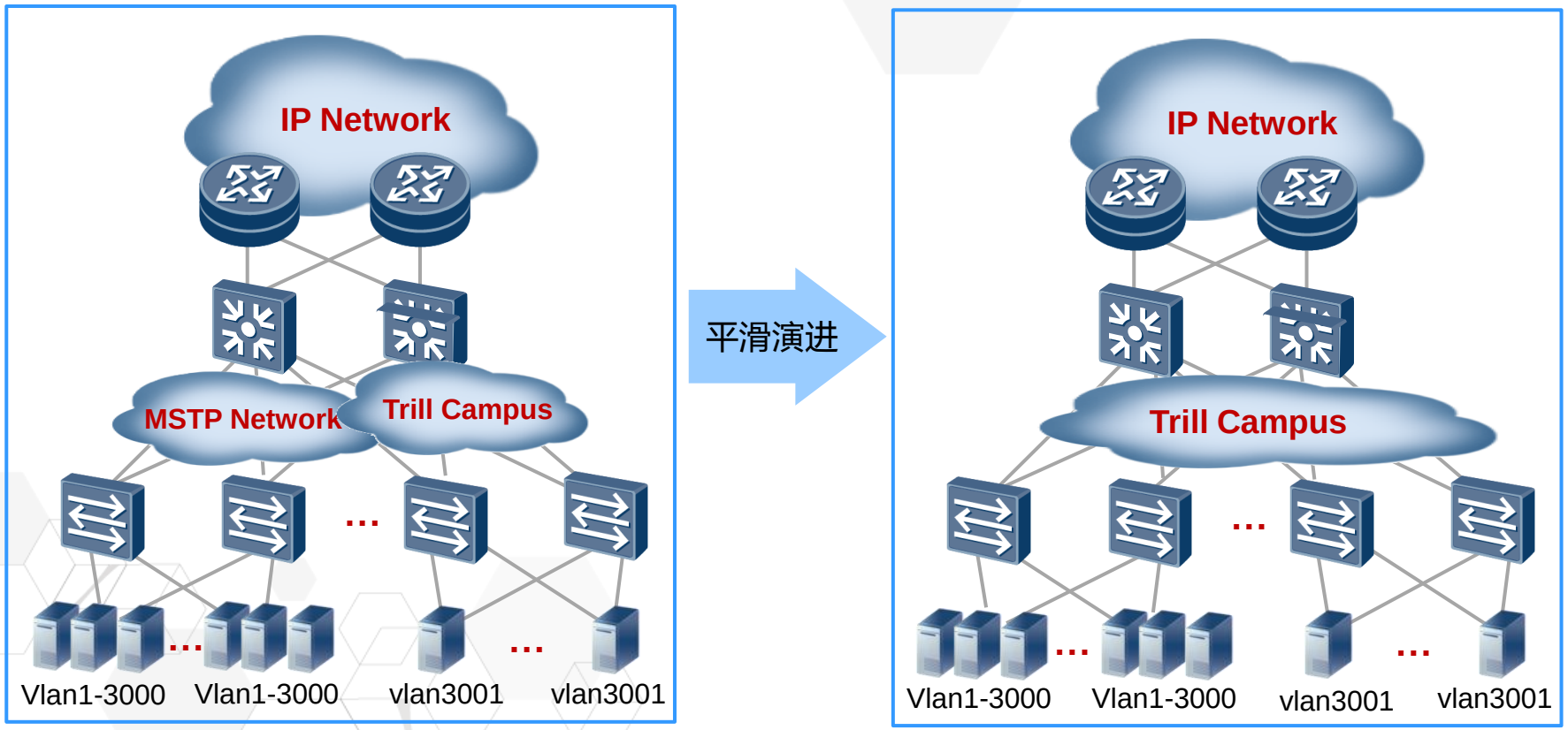
演进方式

- DC初始阶段采用传统MSTP二层网络运营，交换机硬件设备不支持TRILL（比如S9300）。
- 后续DC扩容，新设备支持TRILL转发和大二层组网能力（比如CE12800），可以将原有MSTP网络设备纳入到整个大二层网络中。服务器能够任意挂接在MSTP或TRILL网络的接入服务器下面， VM可以在整个大二层网络下任意迁移

实现方式

- TRILL网络边缘设备模拟MSTP根桥，和MSTP网络进行互通；边缘设备收到MSTP网络拓扑变化TCN报文，在清除自身MAC之后，也能够通知远端RB清除实例相关MAC。

TRILL组网应用—DC平滑演进二



运维方式平滑演进！

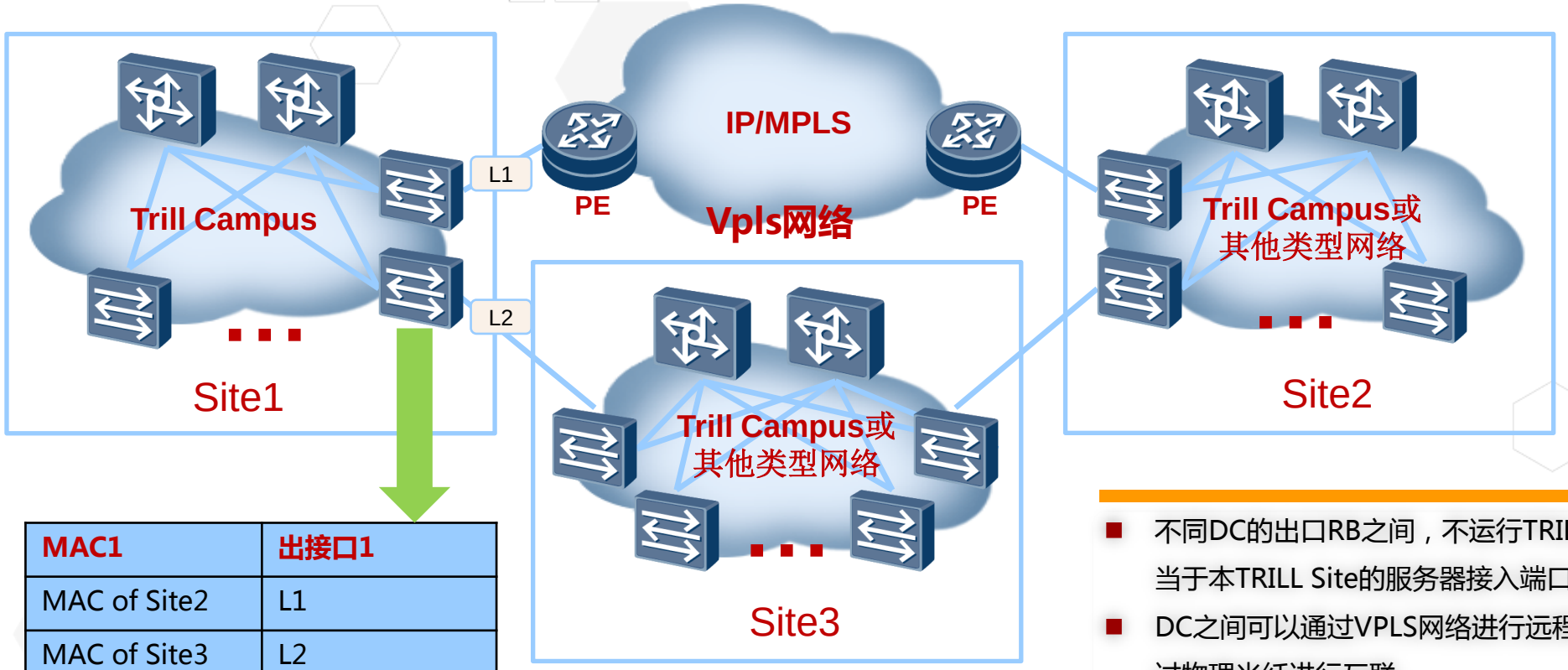
演进方式

- DC全部采用新一代数据中心交换机进行组网，但是由于大二层运营经验不丰富，初始只有少量试点业务（VLAN3001）通过TRILL网络进行运营，其余还是采用传统MSTP进行运营（VLAN 1-3000）。
- 后续随着大二层运维经验的不断积累，所有业务都切换到TRILL网络中，减少整网协议数量，简化运维管理。

实现方式

- 通过灵活指定服务器接入C-VLAN接入的网络类型实现。比如初始指定VLAN 3001接入TRILL网络，后续将VLAN 1-3000也指定接入TRILL网络。

TRILL组网应用—DC互联一

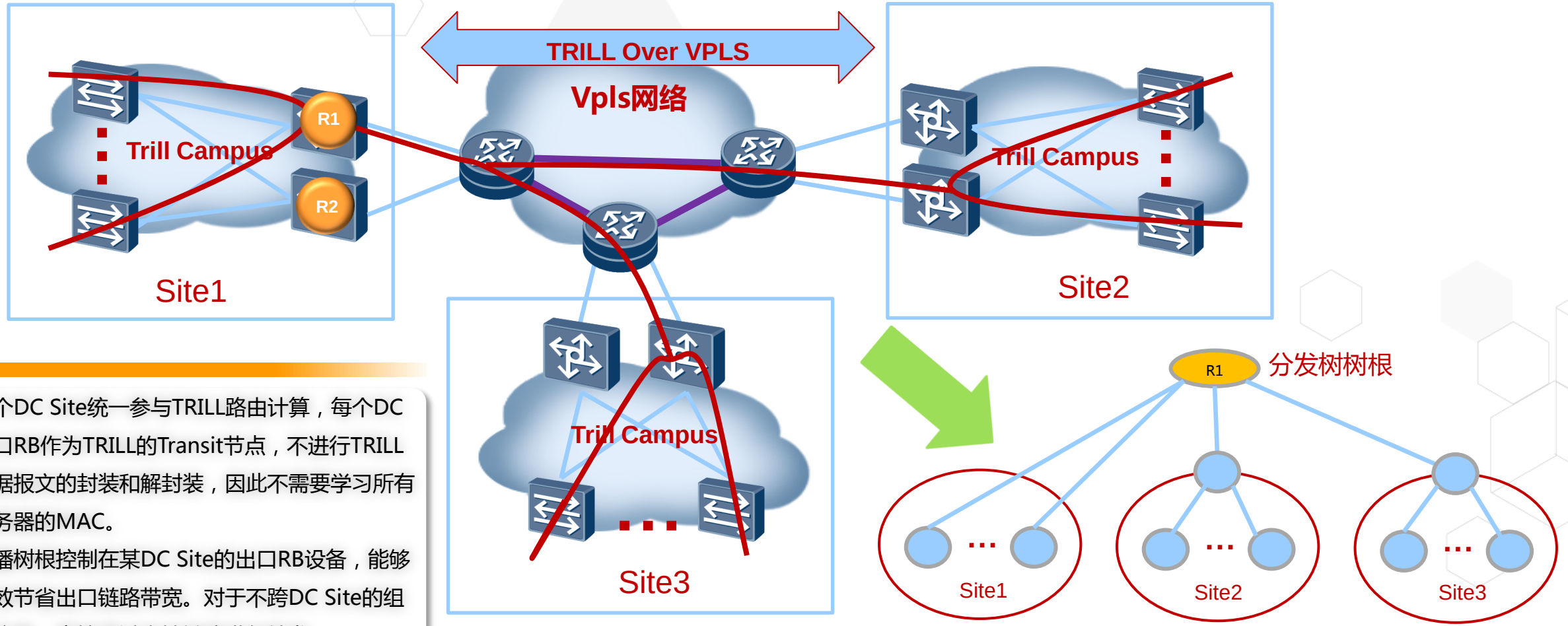


MAC1	出接口1
MAC of Site2	L1
MAC of Site3	L2
MAC of Site1	Site1内部接口

- 不同DC的出口RB之间，不运行TRILL协议。互联端口相当于本TRILL Site的服务器接入端口。
- DC之间可以通过VPLS网络进行远程互联，也可以直接通过物理光纤进行互联。
- 出口RB学习所有服务器的MAC，MAC压力大！

支持TRILL和其他网络的异构互联！

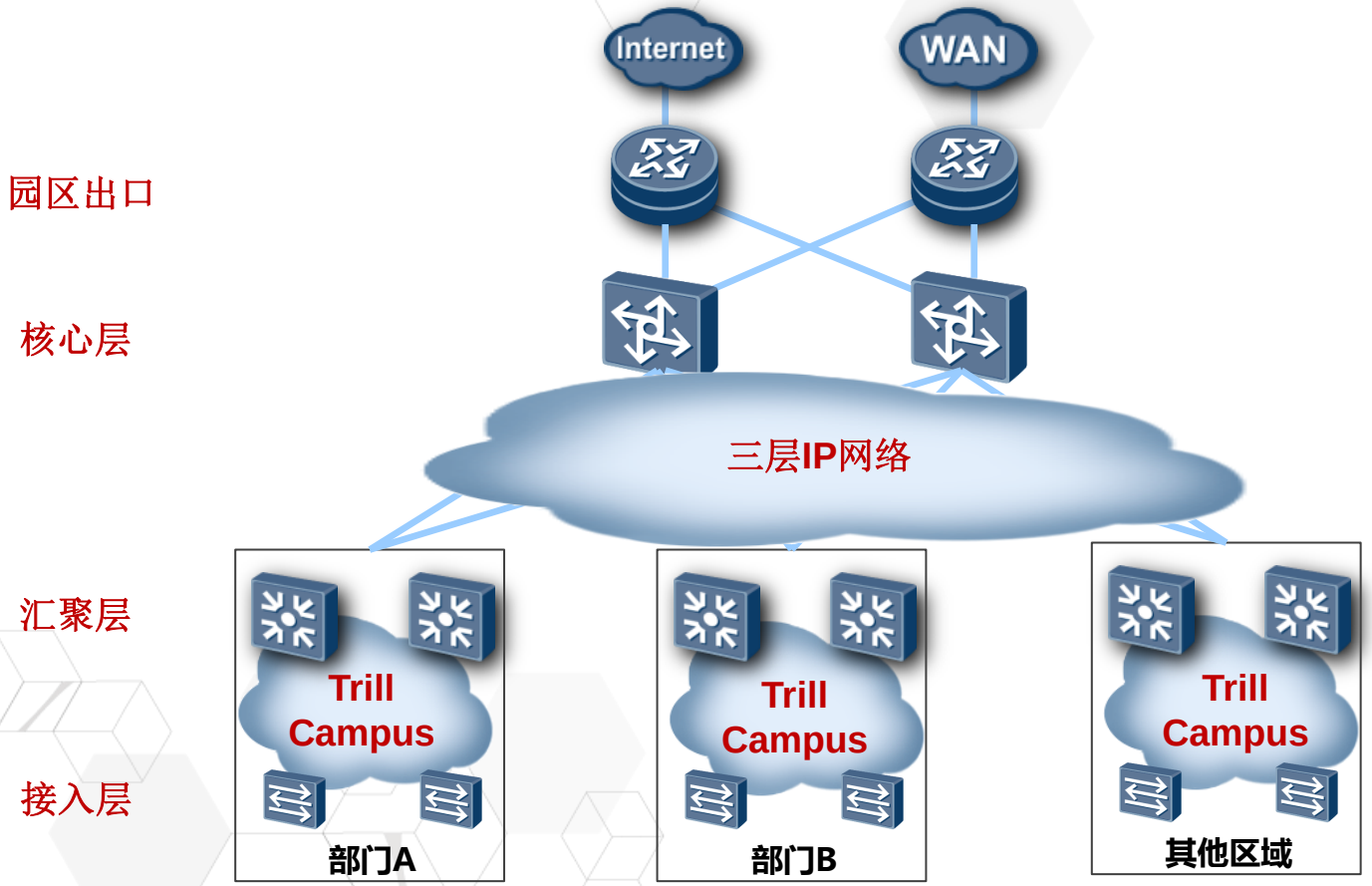
TRILL组网应用—DC互联二



- 多个DC Site统一参与TRILL路由计算，每个DC出口RB作为TRILL的Transit节点，不进行TRILL数据报文的封装和解封装，因此不需要学习所有服务器的MAC。
- 组播树根控制在某DC Site的出口RB设备，能够有效节省出口链路带宽。对于不跨DC Site的组播流量，直接通过本地链路进行转发。

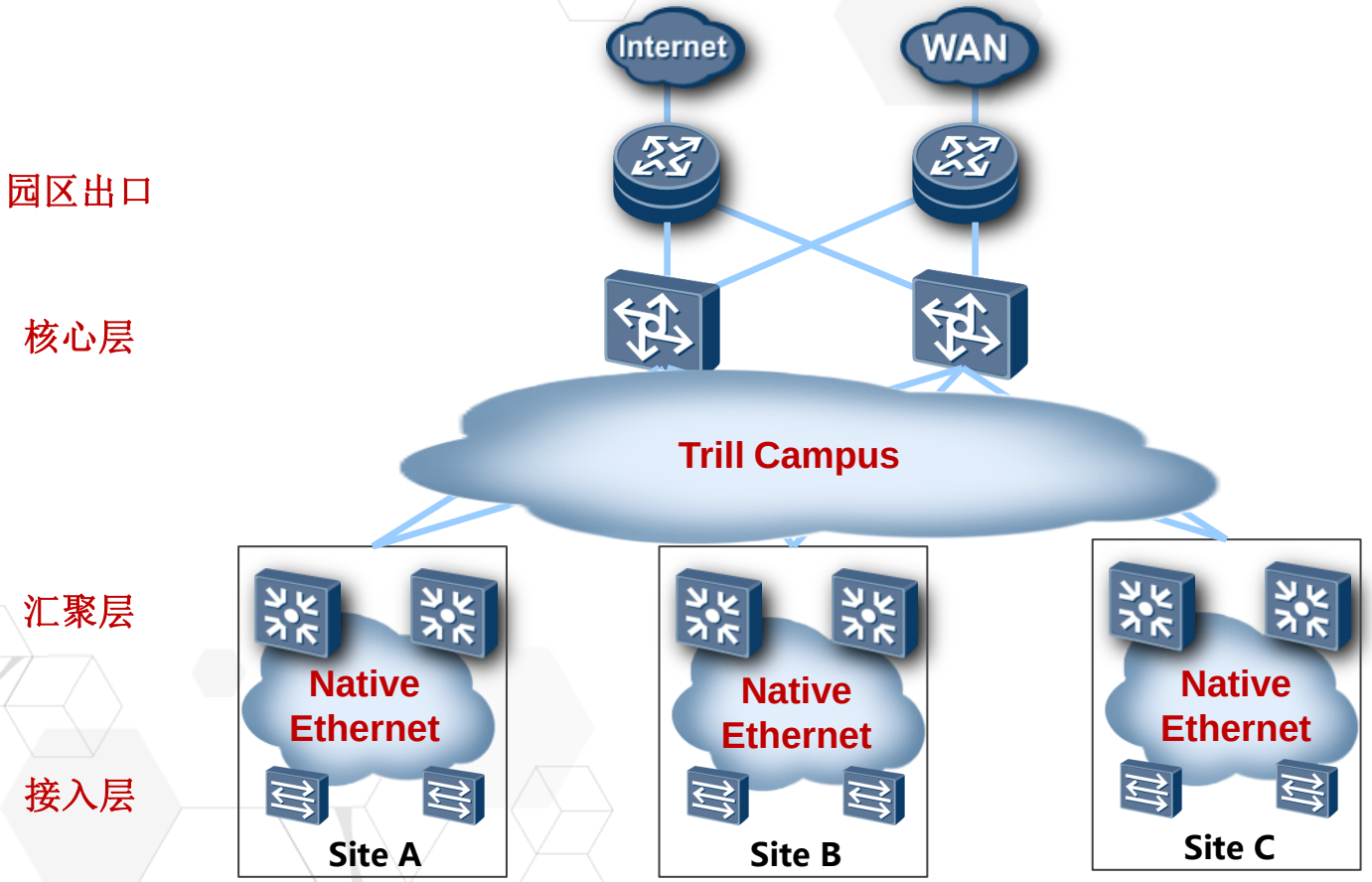
出口RB设备MAC压力小！

TRILL组网应用—园区网应用一



- 接入层和汇聚层：二层组网，可以使用TRILL协议替代传统MSTP组网方式，提升网络带宽利用率、收敛时间以及部署自动化程度，也能够有效避免环路风暴。
- 汇聚层和核心层：部门之间仍然采用三层组网方式，承载部门之间横向流量。
- 汇聚层设备做网关，衔接L2和L3网络。

TRILL组网应用—园区网应用二



- 同一个业务系统分布在不同地理位置（比如不同大楼），业务系统内部需要二层互通。
- 每个Site通过Native ETH传统二层方式组网，Site之间通过Trill进行互联，承载不同Site之间的横向流量，从而构建大二层网络。
- 网关位于园区出口路由器或核心交换机上，衔接L2和L3网络。

缩略语

缩略语	英文	中文
RB	Router Bridge	路由桥接设备，运行TRILL协议的网络设备
AF	Appointed Forwarder	指定转发者
DRB	Desginated Router Bridge	指定RB，在每条运行TRILL Hello的链路上会选举出唯一的RB作为DRB
SPF	Shortest Path First	最短路径优先
IS-IS	Intermediate System to Intermediate System	中间系统-中间系统协议
LSP	Link State PDU	链路状态PDU
P2P	Point-to-point	点对点
VS	Virtual Switch	逻辑路由器
RPF	Reverse Path Forwarding	反向路径转发